ACSIJ

WWW.ACSIJ.ORG

Advances in Computer Science: an International Journal

Vol. 4, Issue 6, November 2015

www.ACSIJ.org

ISSN : 2322-5157

ACSIJ Reviewers Committee 2015

- **Prof. José Santos Reyes**, Faculty of Computer Science, University of A Coruña, Spain
- **Dr. Dariusz Jacek Jakóbczak**, Technical University of Koszalin, Poland
- **Dr. Artis Mednis**, Cyber-Physical Systems Laboratory Institute of Electronics and Computer Science, Latvia
- **Dr. Heinz DOBLER**, University of Applied Sciences Upper Austria, Austria
- **Dr. Ahlem Nabli**, Faculty of sciences of Sfax,Tunisia
- **Prof. Zhong Ji**, School of Electronic Information Engineering, Tianjin University, Tianjin, China
- **Prof. Noura AKNIN**, Abdelmalek Essaadi University, Morocco
- **Dr. Qiang Zhu**, Geosciences Dept., Stony Brook University, United States
- **Dr. Urmila Shrawankar**, G. H. Raisoni College of Engineering, Nagpur, India
- **Dr. Uchechukwu Awada**, Network and Cloud Computing Laboratory, School of Computer Science and Technology, Dalian University of Technology, China
- **Dr. Seyyed Hossein Erfani**, Department of Computer Engineering, Islamic Azad University, Science and Research branch, Tehran, Iran
- **Dr. Nazir Ahmad Suhail**, School of Computer Science and Information Technology, Kampala University, Uganda
- **Dr. Fateme Ghomanjani**, Department of Mathematics, Ferdowsi University Of Mashhad, Iran
- **Dr. Islam Abdul-Azeem Fouad**, Biomedical Technology Department, College of applied Medical Sciences, SALMAN BIN ABDUL-AZIZ University, K.S.A
- **Dr. Zaki Brahmi**, Department of Computer Science, University of Sousse, Tunisia
- **Dr. Mohammad Abu Omar**, Information Systems, Limkokwing University of Creative Technology, Malaysia
- **Dr. Kishori Mohan Konwar**, Department of Microbiology and Immunology, University of British Columbia, Canada
- **Dr. S.Senthilkumar**, School of Computing Science and Engineering, VIT-University, INDIA
- **Dr. Elham Andaroodi**, School of Architecture, University of Tehran, Iran
- **Dr. Shervan Fekri Ershad**, Artificial intelligence, Amin University of Isfahan, Iran
- **Dr. G.UMARANI SRIKANTH**, S.A.ENGINEERING COLLEGE, ANNA UNIVERSTIY, CHENNAI, India
- **Dr. Senlin Liang**, Department of Computer Science, Stony Brook University, USA
- **Dr. Ehsan Mohebi**, Department of Science, Information Technology and Engineering, University of Ballarat, Australia
- **Sr. Mehdi Bahrami**, EECS Department, University of California, Merced, USA
- **Dr. Sandeep Reddivari**, Department of Computer Science and Engineering, Mississippi State University, USA
- **Dr. Chaker Bechir Jebari**, Computer Science and information technology, College of Science, University of Tunis, Tunisia
- **Dr. Javed Anjum Sheikh**, Assistant Professor and Associate Director, Faculty of Computing and IT, University of Gujrat, Pakistan
- **Dr. ANANDAKUMAR.H**, PSG College of Technology (Anna University of Technology), India
- **Dr. Ajit Kumar Shrivastava**, TRUBA Institute of Engg. & I.T, Bhopal, RGPV University, India

ACSIJ Published Papers are Indexed By:

Google Scholar
EZB, Electronic Journals Library (University Library of Regensburg, Germany)
DOAJ, Directory of Open Access Journals
Bielefeld University Library - BASE (Germany)
Academia.edu (San Francisco, CA)
Research Bible (Tokyo, Japan)
Academic Journals Database
Technical University of Applied Sciences (TH - WILDAU Germany)
AcademicKeys
WorldCat (OCLC)
TIB - German National Library of Science and Technology
The University of Hong Kong Libraries
Science Gate
OAJI Open Academic Journals Index. (Russian Federation)
Harvester Systems University of Ruhuna
J. Paul Leonard Library _ San Francisco State University
OALib _ Open Access Library
Université Joseph Fourier _ France
CIVILICA (Iran)
CiteSeerX _ Pennsylvania State University (United States)
The Collection of Computer Science Bibliographies (Germany)
Indiana University (Indiana, United States)
Tsinghua University Library (Beijing, China)
Cite Factor
OAA _ Open Access Articles (Singapore)
Index Copernicus International (Poland)
Scribd
QOAM _ Radboud University Nijmegen (Nijmegen, Netherlands)
Bibliothekssystem Universität Hamburg
The National Science Library, Chinese Academy of Sciences (NSLC)
Universia Holding (Spania)
Technical University of Denmark (Denmark)

TABLE OF CONTENTS

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Modeling and Simulation of Fire Evacuation in Public Buildings

Manh Hung Nguyen[1,2], Tuong Vinh Ho[2] and Jean-Charles Richaud[3]

[1] Posts and telecommunications Institute of Technology (PTIT)
Hanoi, Vietnam

[2] IRD, UMI 209 UMMISCO; IFI/MSI, Vietnam National University in Hanoi
Hanoi, Vietnam

[3] IFI, Vietnam National University in Hanoi
Hanoi, Vietnam

Email: nmhufng@yahoo.com, ho.tuong.vinh@ifi.edu.vn, richaud.p19@ifi.edu.vn

Abstract

The negative consequence of fire, especially fire in public buildings, brings too much of lost in both human and money. The fire evacuation specialists proposed many evacuate techniques, methods and policies adapting to the given building, groups of people, or situations. However, conducting experiments to test these proposed solutions, in the reality, is nearly impossible. Therefore, simulation of fire and fire evacuation to evaluate these proposals is a reasonable solution. This paper proposes an agent-based model for modeling and simulation of fire evacuation in public buildings. The model is implemented and tested using the GAMA agent-based simulation platform.

Keywords: Modeling, simulation, fire evacuation, multiagent system

1. Introduction

The negative consequence of fire, especially fire in public buildings, brings too much of lost in both human and money. The fire evacuation specialists proposed many evacuate techniques, method and policy bay on the given building, groups of people, or situations. However, taking experiments to test these proposed solutions, in the reality, is impossible. Therefore, simulation of fire and fire evacuation to test these proposal is an acceptable solution.

Recently, modeling and simulation of fire evacuation system is one of the most interesting research subjects. Most of proposed models are agent-based modeling and simulation. In which, each agent is autonomy. It could move to other position to meet or interact to other agents to reach its goal. These features of multiagent system are naturally appropriate to the simulation of fire evacuation: an agent could play the role of an evacuee, a fire fighter, a fire evacuation router or some simpler objects such as a fire, a smoke, a water to eliminate fire and smoke, an alarm, an evacuation sign, etc. During fire evacuation, these agents have some actions or behaviors: observe the fire and smoke to avoid them or call the police or to warn others, evacuate themselves or follow other or follow the instructions of fire evacuation routers or policemen, help other to evacuate, help the fire fighter to eliminate fire and smoke, etc. These activities or behaviors could be modeled and realized by using agent technology. That is why most proposed model in the domain is agent-based. For instances, the model of Okaya and Takahashi [1]; Saelao and Patvichaichod [2]; Filippoupolitis [3]; Tang and Ren [4]; Averill and Song [5]; Yi and Shi [6].

In the line with our previous work on modeling and simulation of fire evacuation in public building (Nguyen et al. [7],[8],[9]), this paper proposes an agent-based model for modeling and simulation of fire evacuation in public buildings. The model is implemented in the simulation platform of GAMA (Amouraux et al. [10]).

The paper is organized as follow: Section 2 presents the proposed model. Section 3 presents the allying of the model in a case study. Section 4 is a conclusion and perspectives.

2. Modeling of Fire Evacuation System

2.1 Extension from SEBES model

As mentioned, the model in this paper is extended from the SEBES model (Nguyen et al. [7]). Therefore, the model is called SEBES+.

Inheriting from SEBES model, the SEBES+ has five kinds of agent:

- *fire*: representing fire. The fire agent could propagate within the building space.

- *smoke*: representing smoke. The smoke agent is created from fire agents. It could propagate inside the building space and therefore increase the smoke intensity at a give position by time.
- *alarm*: representing a fire alarm. This agent could detect fire/smoke in its detection range and ring in a ringing duration of time.
- *sign and plan*: representing of evacuation signs and plan. This is a non-movable agent. This provides the information about the direction to emergency exits.
- *evacuee*: representing an evacuee. This agent could see the fire/smoke, hear the alarm, and evacuate to one of the emergency exits by avoiding the obstacles and other evacuees.

The model SEBES+ is added some new kinds of agent for the objective of fire fighting:

- *water/steam*: representing water or water steam to eliminate fire and smoke. This agent is borne from FireFighter water sources or extinguisher.
- *extinguisher*: representing extinguisher device. This agent is put inside the building and its position is noted in evacuation plan. It is used by FireFighter to generate carbon-dioxide to eliminate fire and smoke.
- *fire fighter*: representing fire fighter. This agent play multi role during fire evacuation: fire fighter to bring water and/or extinguisher to eliminate fire and smoke; evacuation router to rout evacuees to move to emergency exits; and act as an evacuee.

The next section will present these new extended agents in detail.

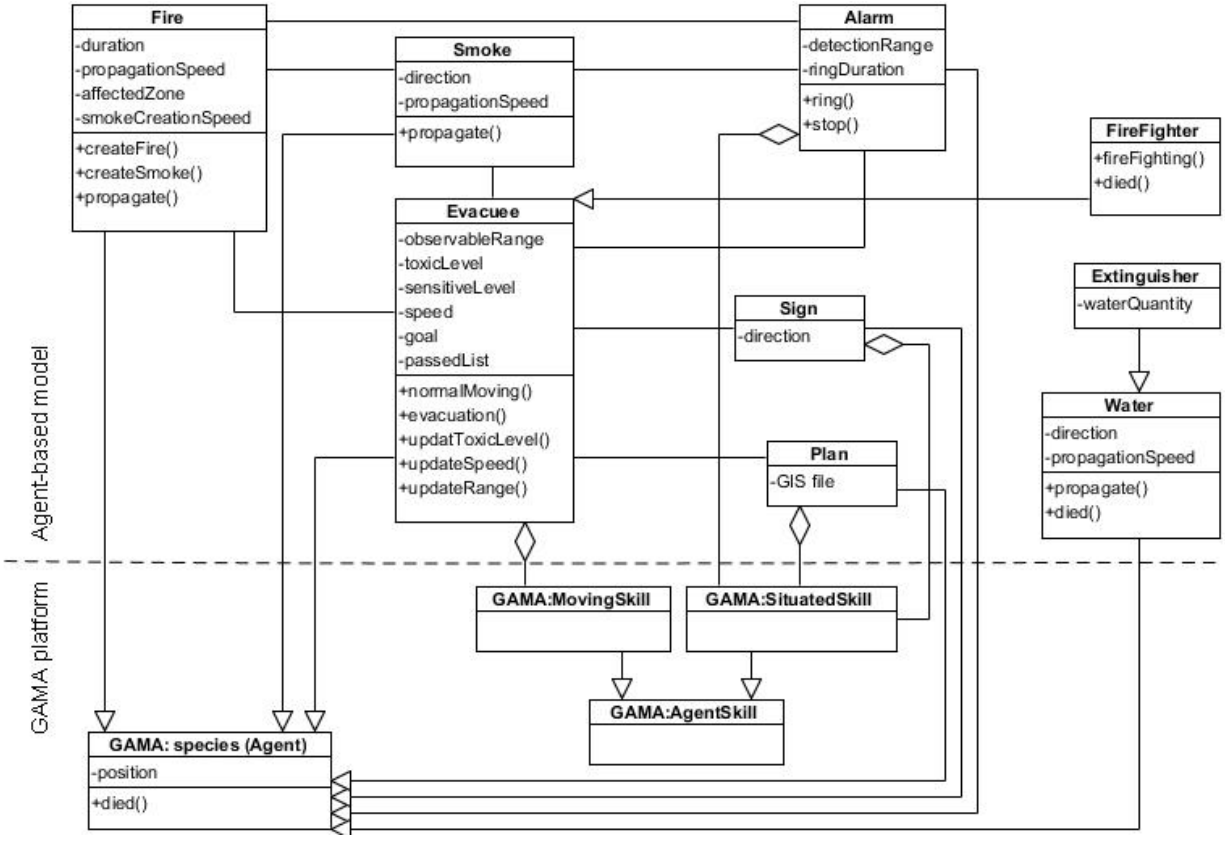


Fig. 1 Extended class diagram of SEBES+ from SEBES model.

.

2.2 Water agent

2.2.1 Attributes

- *power*: the ability to eliminate fire and smoke of it.
- *range*: the effected zone of it. It could eliminate fire and/or smoke if these agents go inside its zone.
- *direction*: the direction of propagation.
- *propagationSpeed*: the speed to propagate of this agent.

2.2.2 Actions

- *propagate*: it propagates in its direction inside the building.
- *died*: it will be died after eliminate fire and/or smoke. The elimination rate is 2:1: two water agents could eliminate a fore or a smoke agent.

2.3 Extinguisher agent

2.3.1 Attributes

- *waterQuantity*: the amount of water or carbon-dioxide which could be generated by this agent when used. This could be recharged by fire technician.

2.3.2 Actions

- *propagate*: it propagates in the direction made by the fire fighter.
- *died*: it will be died after using all water quantity.

2.4 FireFighter agent

2.4.1 Attributes

This agent is inherited from *evacuee* agent, so it has all attributes and activities of that agent. It also has some new added attributes:

- *objective*: the object of the fire fighter during fire evacuation. It could be assigned a value of: evacuating, fire fighting, people routing, people helping.

2.4.2 Actions

Beside activities extended from *evacuee* agent, this agent has some new added actions:

- *fireFighting*: this includes several actions which could be executed in any order: moving to extinguishers position to get it to use, using extinguisher to eliminate fire and/or smoke, routing people to evacuate as quick as possible to the emergency exits, helping people to safety evacuate.

3. Case Study

3.1 Choice of Public Building

In our case study, we apply the proposed model for the building that Préventex used in their case study.

Préventex was the first joint sector-based association created in the private sector under the Act respecting occupational health and safety. On October 22, 1981, the Board of Directors of the Commission de la santé et de la sécurité du travail (CSST - the Quebec Health and Safety Commission) adopted resolution A-170-81 that was to lead to the creation of the joint sector-based association of health and safety for the textile and knitting industry.

The evacuation plan of the building is presented in the Figure 2. It is composed of three zones (Source : http://www.preventex.qc.ca/images/ documents/info/en/evacuation.pdf):

- Zone A: This zone is composed of a coffee room , a waiting room, and an emergency door. There are five extinguishers in total in this zone.
- Zone B: This zone is composed of a reception room, a coffee room, stairs doors and enter door. There are totally five extinguishers and three emergency doors in this zone.
- Zone C: This zone is composed of an office room, WC, and two laboratories. There are totally six extinguishers and three emergency doors in this zone.

3.2 Choice of Simulation Platform

On the platform of the simulation, the model is developed on the simulation platform GAMA (Amouroux et al. 2007). GAMA provides a simulation development environment for building spatially explicit agent-based simulations. It enables: (i) to use arbitrarily complex GIS data as environments for the agents; (ii) to run simulations composed of vast numbers of agents; (iii) to conduct automated controlled experiments on various scenarios, with a systematic, guided or ‘‘intelligent’’ exploration of the space of parameters of models; and (iv) to let users interact with the agents in the course of the simulations.

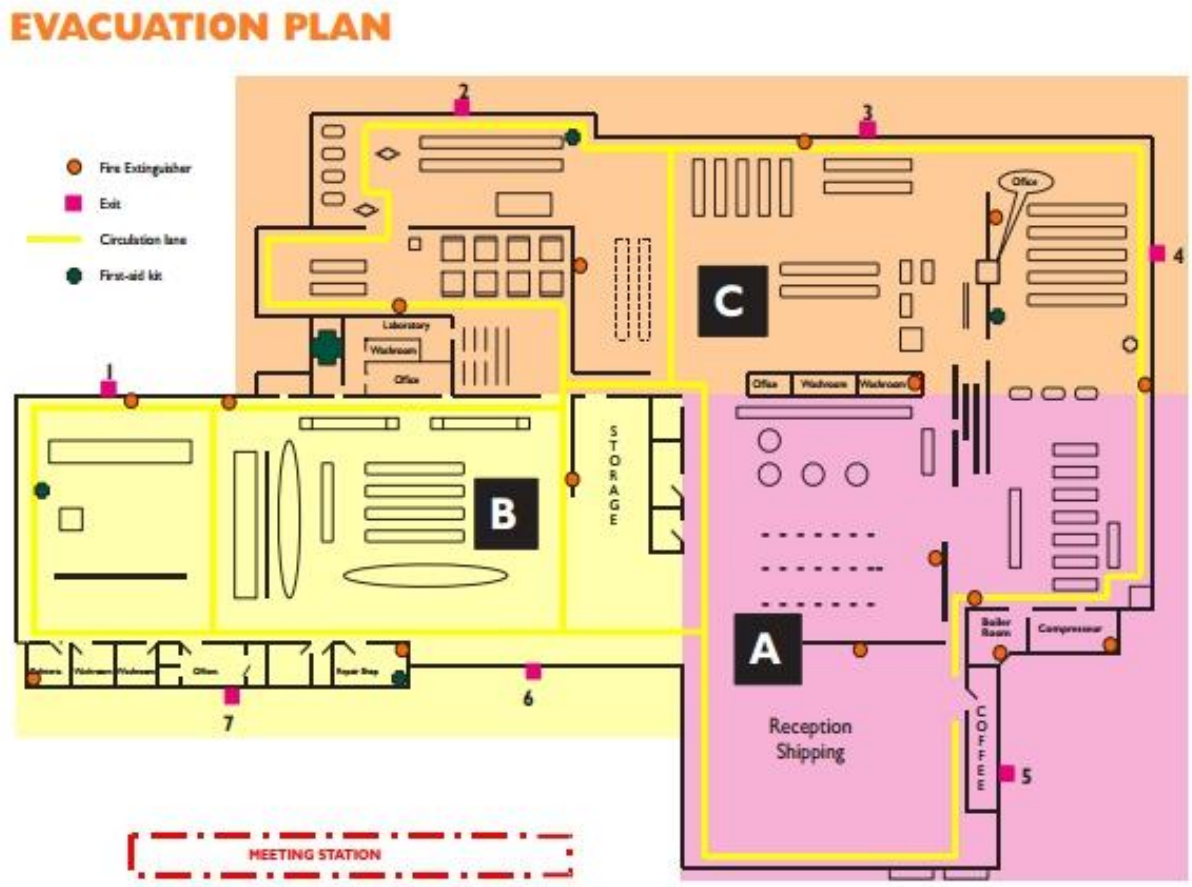


Fig. 2 Evacuation plan of the building based on Préventex's case study.

3.3 Results

3.3.1 Fire and smoke propagation in fighting against water

The propagation of fire/smoke against water/steam is visually presented in Figure 2. In which, fire/smoke agent is presented in gray cycles. Water/steam agent is presented in blue rectangles. The fighting rate of water:smoke is 2:1. It means that two water agents could kill a smoke agent.
The visualization results indicates that the smoke agent number is reduced near the position of water/steam agents.

3.3.2 Evolution of fire and smoke propagation

The evolution of fire/smoke against water/steam is presented in Figure 3: fire/smoke is represented in gray line; and water/steam is represented in blue line.
This simulation indicates that the fire/smoke occurs at the time 100 and grows up gradually. When the fire evacuation system detected the fire in the building, it kicks off the alarm system and the firefighter arrives. They start to puff of water/steam into fire/smoke (at the time 150), and then, the quantity of fire/smoke decreased when the number of water/steam increases (at the time 200). This tendency continues until the fire/smoke if totally eliminated and the firefighters stop their mission at the fire site.

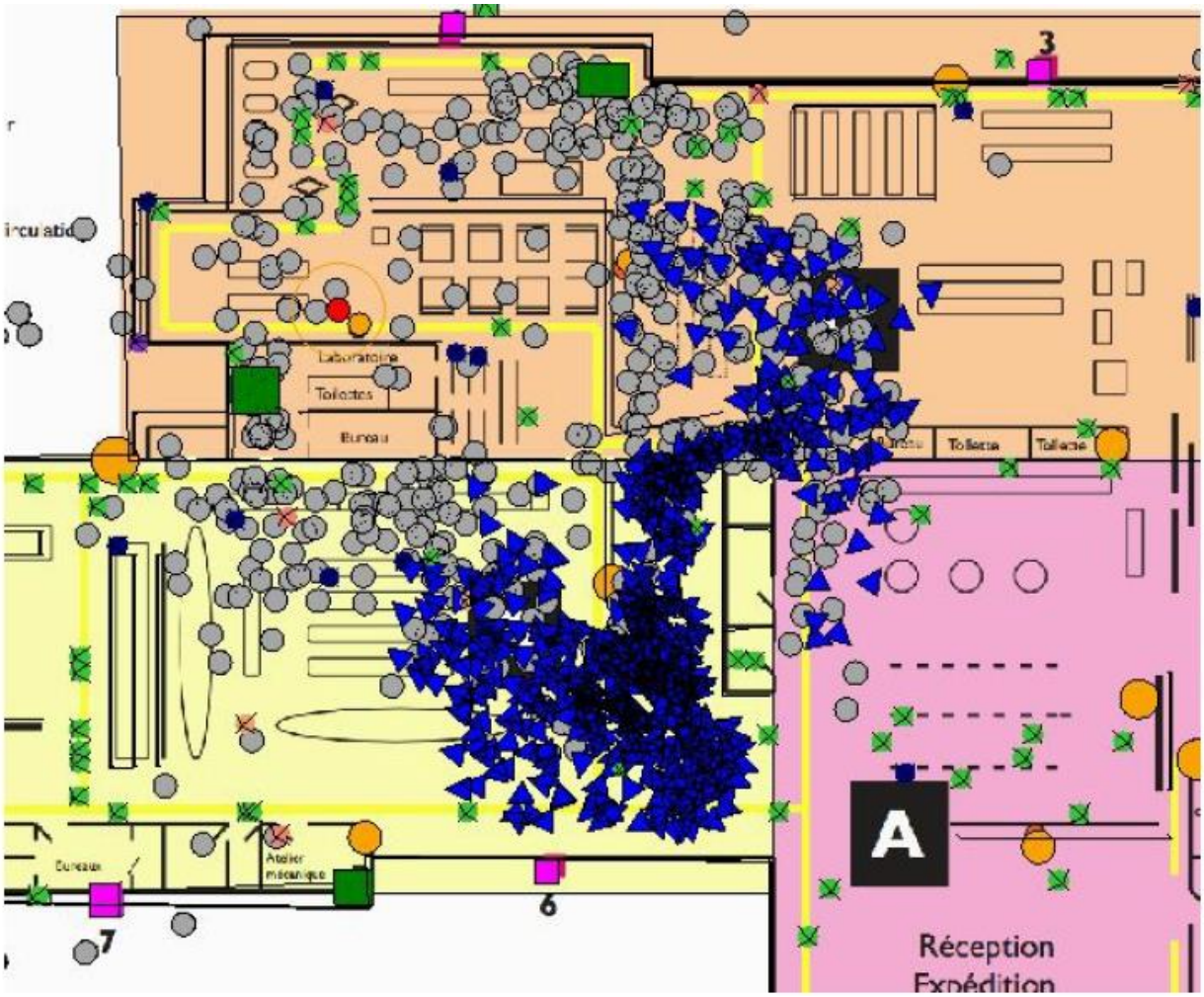


Fig. 3 Visualization of fire and smoke propagation against water.

3.3.3 Evolution of people objective during fire evacuation

We distinguish four kinds of people objective (or activity):

- Working: they are working (black line).
- Circulation lane: They are in the circulation lane to evacuate (yellow line).
- Emergency exit: They arrive at one of emergency exit. (matron line).
- Meeting station: They are safety at the meeting station (red line).

The evolution of the number of people in each activity is represented in Figure 5. At the beginning time (0-200), most of people is in their work. When the fire occurs and the alarm rings, people changes their behavior: they start to regroup in circulation lane to evacuate. So the number of people in circulation line increases. The number of people at emergency exits is proportionally related to the number of people in the circulation lines at previous time. Consequently, the number of escaped people (at the meeting station) increases until there is no more people inside the fire building.

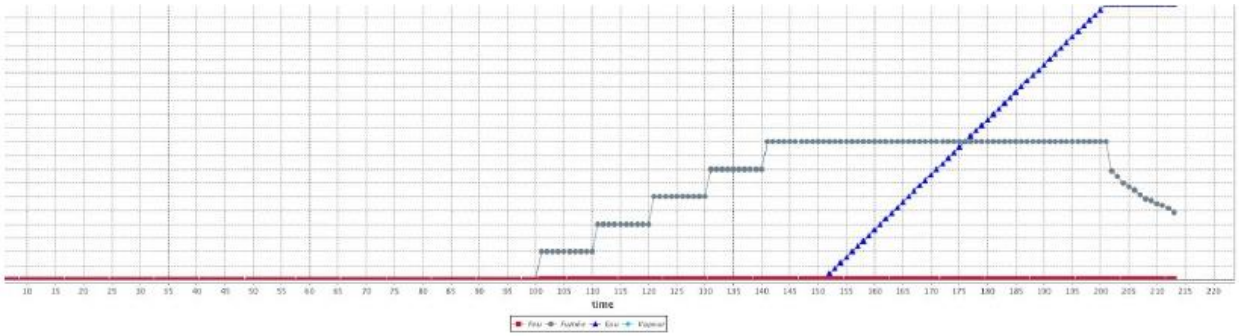


Fig. 4 Evolution of fire and smoke propagation.

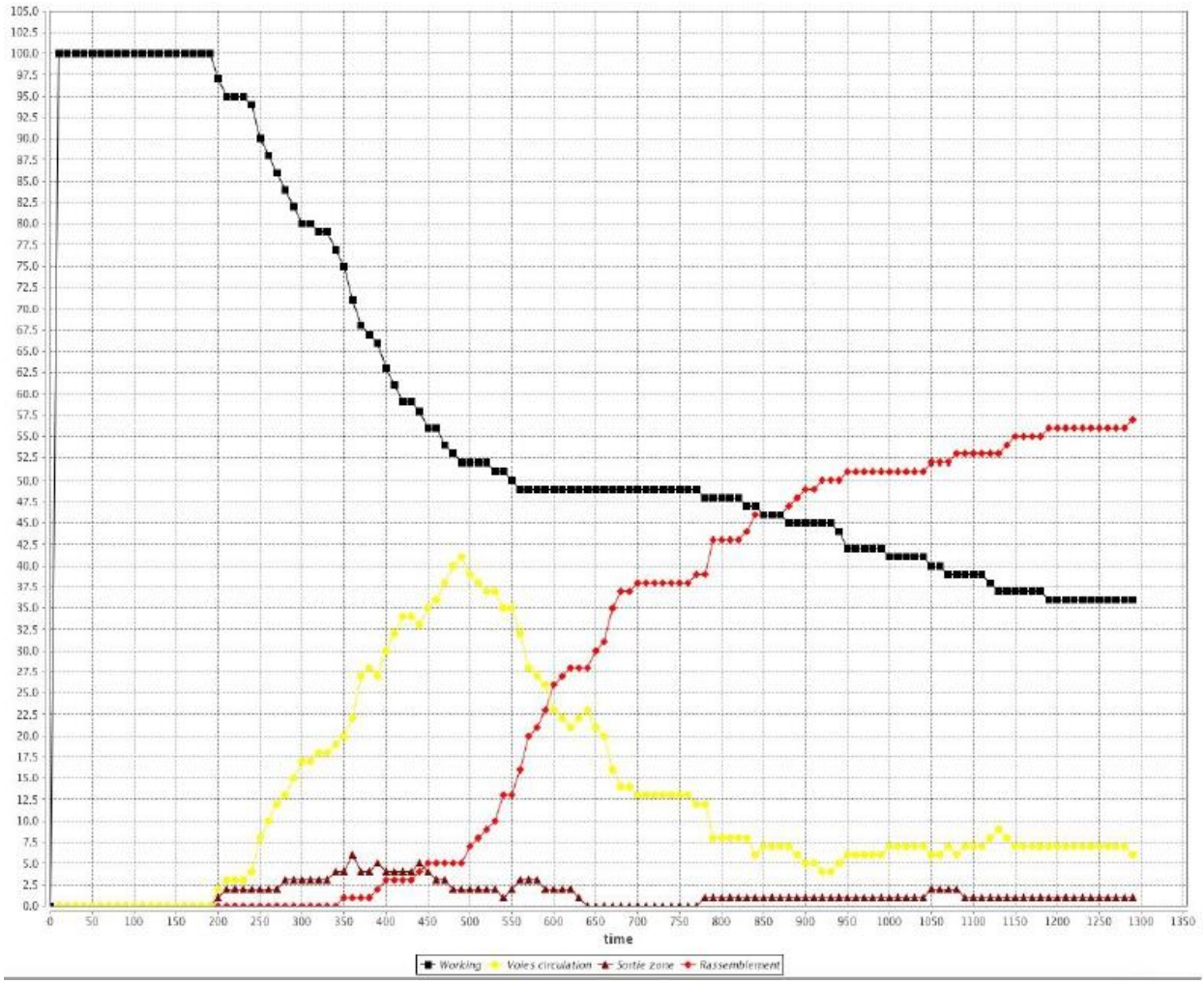


Fig. 5 Evolution of people objective during fire evacuation.

4. Conclusions

This paper presented an agent-based model for modeling and simulation of fire evacuation inside public building, called SEBES+ model. This model is an extension from the SEBES model by adding some new kinds of agent in the system: water or water steam agent to eliminate fire and/or smoke, extinguisher agent to generate water or carbon-dioxide to eliminate fire, fire fighter agent to use water and/or extinguisher to eliminate fire/smoke, to rout and help people in evacuating, etc.

The proposed model has been implemented in GAMA, an agent-based simulation platform and then, tested with some scenarios.

The validation and evaluation of the proposed model, and the testing of some new fire evacuation policy are some of our perspective works in the near future.

Acknowledgments

This work is funded by the research project at Vietnam National University in Hanoi, number QG.15.31, on the modeling and simulation of fire evacuation in public buildings.

References

[1] M. Okaya, T. Takahashi, Human relationship modeling in agentbased crowd evacuation simulation, in: D. Kinny, J.Y. jen Hsu, G. Governatori, A.K.Ghose (Eds.), Proceedings of Agents in Principle, Agents in Practice – 14th International Conference, PRIMA 2011, Wollongong, Australia, November 16–18, 2011, Lecture Notes in Computer Science, vol. 7047, Springer, 2011, pp. 496–507.

[2] T. Saelao, S. Patvichaichod, The computational fluid dynamic simulation of fire evacuation from the student dormitory, American Journal of Applied Sciences 9 (3) (2012) 429–435.

[3] A. Filippoupolitis, An adaptive system for movement decision support in building evacuation, in: Proceedings of the 25th Int. Symposium on Computer and Information Sciences, London, UK, 2010, pp. 389–392.

[4] F. Tang, A. Ren, Agent-based evacuation model incorporating fire scene and building geometry, Tsinghua Science Technology 13 (5) (2008) 708–714.

[5] J.D. Averill, W. Song, Accounting for Emergency Response in Building Evacuation: Modeling Differential Egress Capacity Solutions, 2007.

[6] S. Yi, J. Shi, An agent-based simulation model for occupant evacuation under fire conditions, in: Proceedings of the 2009 WRI Global Congress on Intelligent Systems, GCIS '09, vol. 01, IEEE Computer Society, Washington, DC, USA, 2009, pp. 27–31.

[7] Manh Hung Nguyen, Tuong Vinh Ho and JeanDaniel Zucker. Integration of Smoke Effect and Blind Evacuation Strategy (SEBES) within Fire Evacuation Simulation. Simulation Modelling Practice and Theory. Volume 36, August 2013, p.44-59, ISSN 1569-190X.

[8] Manh Hung Nguyen, Tuong Vinh Ho and Jean-Daniel Zucker. A Simulation Model for Optimise the Fire Evacuation Configuration in the Metro supermarket of Hanoi. Proceedings of the Ninth International Conference on Simulated Evolution And Learning (SEAL2012), Hanoi, Vietnam, 16-19 December 2012. L.T. Bui et al. (Eds.): SEAL 2012, LNCS 7673, pp. 470–479, Springer-Verlag Berlin Heidelberg 2012.

[9] Manh Hung Nguyen, Tuong Vinh Ho, Thi Ngoc Anh Nguyen and Jean-Daniel Zucker. Which Behavior is best in a Fire Evacuation: Simulation with the Metro supermarket of Hanoi. Proceedings of The 9th IEEE - RIVF International Conference on Computing and Communication Technology. Ho Chi Minh city, Viet Nam, p.183 -- 188, February 27 - March 1, 2012.

[10] E. Amouroux, C. Quang, A. Boucher, A. Drogoul, GAMA: an environment for implementing and running spatially explicit multi-agent simulations, in: 10th Pacific Rim International Workshop on Multi-Agents (PRIMA), Thailand, 2007.

Manh Hung Nguyen received his Bachelor degree of Information Technology (IT) at The Posts and Telecommunication Institute of Technology (PTIT) in 2004, his Master degree in IT at the L'Institut de la Francophonie pour l'Informatique (old IFI) in 2007, and his Ph.D in IT at the University of Toulouse III Paul Sabatier, France, in 2010. He is currently a lecturer at The Posts and Telecommunication Institute of Technology (PTIT), Hanoi, Vietnam. His domains of interest are: Artificial Intelligence, Multiagent system, Modeling and simulation of complex system, Distributed intelligent computing.

Tuong Vinh Ho is a researcher-lecturer at Institute Francophone International (IFI) since 2000. Currently, he is Vice-Director of IFI in charge of research activities, and Head of MSI (Computational Modeling & Simulation of Complex Systems) research team at IFI. He holds a Ph.D. degree in Computer Engineering from the École Polytechnique de Montréal (1999). During 1998-2000, he was a postdoctoral research fellow at the Software Engineering Management Research Laboratory (Université du Québec à Montréal- UQAM). His research interests include Software Engineering and Computational Modeling and Simulation of Complex Systems.

Jean-Charles Richaud is currently a master student at Institute Francophone International (IFI). He is interested in simulation of complex system based on multiagent system, particularly on the platform GAMA.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A Secure and Efficient Routing Protocol with Genetic Algorithm in Mobile Ad-hoc Networks

Atieh Moghaddam[1], Ali Payandeh[2]

[1] Computer Department, University of Tehran, Company
Tehran, IR.TE, Iran
a.moghaddam@alumni.ut.ac.ir

[2] ICT Department, Malek-e-AshtarUniversity, Company
Tehran, IR.TE, Iran
payandeh@mut.ac.ir

Abstract

Routing in Mobile Ad-Hoc Networks (MANETs) is a challenging task due to its nature of open medium, infrastructure-lessness, dynamicity and no trusted central authority. In MANET, a node can be compromised during the route discovery process. Attackers from inside or outside can easily exploit the network. Several secure routing protocols have been proposed for MANETs. In this paper, Ad-Hoc On-Demand Distance Vector (AODV) routing protocol is considered due to the fact that it uses the shortest number of wireless hops towards a destination as the primary metric for selecting a route with independence of the traffic congestion. To add security to AODV, Secure AODV was designed to enhance security services to the original AODV. Secure AODV protocol has been designed with cryptographic techniques such as digital signatures and hash chains, which can have a significant impact on the routing performance of AODV routing protocol. To improve efficiency of SAODV, Enhanced SAODV (ESAODV) was proposed based on Genetic Algorithm and alternative path. The genetic algorithm optimizes the routes in terms of selected metrics. The performance and impacts of using AODV, S-AODV and ESAODV routing protocols were compared using NS-2 Simulator. The simulation results demonstrated that using the proposed mechanism could significantly decrease the End-to-end delay and routing overhead.

Keywords: *Mobile ad-hoc network, SAODV routing protocol, genetic algorithm, end-to-end delay, packet routing overhead.*

1. Introduction

Mobile ad-hoc network is a collection of nodes that are connected to each other with wireless links without any infrastructure. The routing protocols in ad-hoc environment can be classified as proactive routing protocols and reactive routing protocols. Proactive protocols maintain whole paths in routing tables and when the source node wants to establish a route to the destination node, the path that already exists in its routing table is used. In reactive protocols, the route is established only when the source node needs to send a data packet to the destination. There are varieties of routing protocols for MANET such as AODV, DSR, OLSR …, but none of them are secure. As a result, they assume there is no malicious node in the network; however, due to the flexibility of the MANET, there are a lot of vulnerabilities in this kind of network and security problem is the most significant issue in it. Two different security mechanisms are presented for routing protocols. The first one guarantees authentication and integrity of the routing messages. The second mechanism allows node to control another node behavior during route discovery process. Both two approaches need some network resources such as battery, energy and bandwidth. The main purpose is finding the balance between efficiency and security.

The remaining part of this article is organized as follows. Section 2 describes SAODV routing protocol. Section 3 introduces genetic algorithm briefly. A proposed routing protocol is given in section 4. And section 5 shows the simulation results.

2. SAODV Routing Protocol

The first secure and promoted version of AODV is secure AODV (SAODV) that is based on asymmetric cryptography. In SAODV protocol, the routing messages (RREQ, RREP, and RERR) are encrypted by digital signature to guarantee the integrity and authenticity. Due to not propagating the RREQ for external nodes, this routing protocol prevents from external active attacks. All nodes are authenticated by a unique password. When a source node wants to send the RREQ, it first authenticates its neighbors by that password and then broadcasts the message. In SAODV, the sender signs the routing

messages by its private key and the receiver verifies them by the sender's public key. Because of incrementing the hop-count in each step of routing discovery, the sender cannot encrypt it. Hence, for securing this field (that is, not allowing malicious node to reduce it), SAODV uses hash chain.

This structure is difficult to use when an intermediate node has a path to destination in routing table since RREP necessarily has to have destination signature. For solving this problem, SAODV uses double signature. In this mechanism, RREQ has a second signature that is always stored with the reverse path route. An intermediate node, which wants to reply RREQ, uses second signature and adds it to RREP. Then it is sent to the source node. The RREQ and RREP messages fields are:

<Type, Length, Hash function, Max-hop-count, Top Hash, Signature, hash>

The RERR message fields:

<Type, Length, reverse, Signature>

When a node creates RREQ or RERR, It does the functions as follow:

1. Generate a random number (seed)
2. Max-hop-count = timeToLive
3. Hash = seed
4. Hash-function = h
5. Top-hash = $h^{max\text{-}hop\text{-}count}$ (seed)

Verifying hop-count in RREQ or RREP by intermediate node:

1. Top-hash = $h^{max\text{-}hop\text{-}count}$ (Hash)

Update hop-count, apply hash to generate new hash chain, then send it to all neighbors.

However, SAODV messages are significantly larger and require heavy computation because of digital signature, especially for double signature.

SAODV solves the overhead of routing tables by updating them in particular time. So SAODV prevents the black hole attack. In comparison with AODV, due to an encryption in SAODV, malicious node cannot access the content of the messages; nevertheless cryptography process increases routing delay and the length of the messages.

3. Genetic Algorithm

Finding the shortest path in mobile ad-hoc networks requires the evaluation of route from the source to the destination, which has the least cost. The old algorithms such as Dijkstra and Bellman ford present how the shortest path is found. Yet, these algorithms are especially used for wired network and are not suitable for wireless networks. Genetic algorithm is one of the algorithms that are useful for ad-hoc networks and it is used for designing more effective protocols.

John Holland proposed genetic algorithm in 1970. The route consists of sequence of nodes. This algorithm executes on routes, which are achieved from route discovery process. In the first step, the path is coded by sequence of integers, which these are the node's IP. The length of this sequence cannot be more than the number of nodes. GA operation consists of six necessary levels: genetic presentation, initial population, fitness function, selection, crossover and mutation. This collection is "standard GA" (SGA).

3.1 Genetic Presentation

The path is coded by sequence of integers, which are node's IP.

3.2 Initial Population

Each chromosome shows a potential solution. Initial population consists of numbers that represents the chromosomes in AODV protocol. The routes that are achieved from route discovery process are intended for initial chromosomes.

3.3 Fitness Function

The quality of each solution should be evaluated accurately. In this function, the main purpose is finding the richest path between source and destination. The fitness parameters are described according to the problem requirements. In this paper, the goal of using genetic algorithm is finding the shorter path from the source node to the destination node with reducing end-to-end delay.

3.4 Selection

This function plays the significant role to promote the average of population quality by selecting the high-qualified chromosome for next generations. Selection is operated on fitness output. Each chromosome that has the best fitness value is selected. This function plays the

significant role to promote the average of population quality by selecting the high-qualified chromosome for next generations. Selection is operated on fitness output. Each chromosome that has the best fitness value is selected.

3.5 Crossover

Crossover processes the current solutions to find the better approach. In this process, one or more than a bit of chromosome changes and a new population is created. Genes are selected from father's chromosomes and make the new children.

3.6 Mutation

GA could fast access the demanded level of cost. Mutation randomly changes some bits of sequences and move them to new location of existing solution.

4. Proposed Protocol

In this paper, ESAODV is proposed to improve the efficiency of SAODV. ESAODV eliminates the same routing messages, uses genetic algorithm to find the better path in route discovery and also saves an alternative path and uses it when the link failure occurs.

To preserve the security in this protocol, similar to SAODV, digital signature and hash chain are used. This mechanism prevents ESAODV from external attack, eavesdropping and black hole attack.

4.1 Propagation RREQ and RREP

When the source node needs a route to the destination, it creates RREQ and broadcasts it to all neighbors. Intermediate node receives the RREQ packet and then checks the routing table. If there is a route to the destination with higher sequence number, this intermediate node sends the RREP to the source by reverse path. Otherwise, each intermediate node updates its routing table and then sends the RREQ to all neighbors until the destination receives the message.

During the execution of this process, some nodes may give the same RREQ many times and broadcast it more than once, which reduces the energy of nodes and increments the delay. The new mechanism has been designed in ESAODV to prevent from responding the same routing message. When the node receives the RREQ for the first time, it saves its broadcast IP in the routing table. After that, whenever it receives the RREQ with the same IP, it does not broadcast this message because it is reiterative. This solution causes reduction of routing delay and saves the energy of nodes.

As shown in figure 1, A is a source node and 1, 2, 3, 4 and 5 are intermediate nodes. A Broadcasts RREQ to all neighbors and it continues by others. Node 3 is a neighbor of 1 and 2. So it receives RREQ from both 1 and 2 and broadcasts the same RREQ twice.

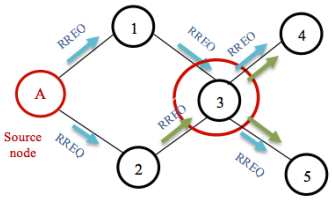


Fig. 1 Broadcasting RREQ in SAODV.

In ESAODV, when node 3 receives the RREQ for the first time, it saves its broadcast IP. After that, whenever it receives the RREQ, firstly it checks the routing table. If the current IP is similar to the IP that exists in routing table, the node eliminates the same RREQ and does not broadcast it. Otherwise, the message is broadcast to its neighbors.

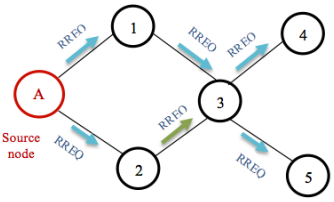


Fig. 2 Broadcasting RREQ in ESAODV.

In the proposed protocol, this structure has also been implemented for broadcasting the RREP. B is the destination node and 4, 5, 6, 7 and 8 are the intermediate nodes.

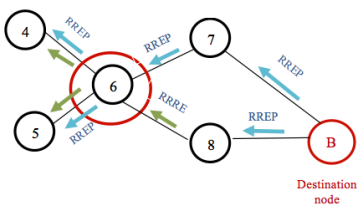


Fig. 3 Broadcasting RREP in SAODV.

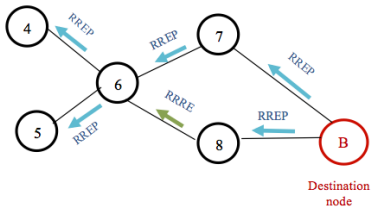


Fig. 4 Broadcasting RREP in ESAODV.

4.2 Implementing the Genetic Algorithm

In the routing process, finding a route, which has less delay, is an important challenge. Smart algorithm is a kind of algorithm that is used in optimization problems to find the better solution. Genetic is one of the smart algorithms that evaluate the chromosome according to the purpose.

In ESAODV, genetic algorithm is executed after the route discovery process. The routes that are found from this process create initial population. Fitness of this algorithm is calculated based on the delay. Each RREQ message has timestamp, which shows the time of the message creation. Route delay is the difference between routing current time (the time that message is received by the destination) and RREQ timestamp.

$$rdelay = (CURRENT\_TIME - rq\_timestamp) \quad (1)$$

rdelay is a variable that shows the delay. Current_time is a time that the message is received by the destination. rq_timestamp is a time that the RREQ has been created. According to the calculated delay for each path, the path that has the least delay is selected. So not only is the rout selected based on the hop-count, but also delay affects selecting it. Genetic algorithm both speeds the routing process and finds the better route to send the data. Due to the reduction of the delay, the network lifetime is increased.

4.3 Alternative Path

After executing route discovery and genetic algorithm, the output of algorithm is selected as a current route to send the information. If link failure occurs, route discovery process in SAODV begins again. This approach increases the packet routing overhead.

ESAODV uses an alternative path. In this mechanism, the second path is saved in routing table of the nodes. When the genetic algorithm is done and the better path is selected for sending the data, the second better path (which has the least delay except the first rout) is selected as an alternative path. So when the link failure occurs, the second path alternates the current route and sending the information continues. To implement this mechanism, each node has rt-count function. This function shows the number of the path to the destination. If this number is less than one, the second path is saved as an alternative path. During the execution of the protocol, if the route with less delay is found, this route is changed with the alternative path and the routing tables are updated. This approach significantly minimizes the packet routing overhead.

5. Simulation Results

NS2 simulator is used to illustrate the performance of the proposed protocol. In this simulation, AODV, SAODV and ESAODV are compared. Table 1 shows the simulation parameters.

Table 1: Simulation parameters

Simulation Time	120 s
Simulation Area Size	1000 * 1000 m
Number of Nodes	30
Data Transfer Rate	4 Packet /s
Wireless nodes transfer scope	250 m
Data packet size	512 bytes
Mobility mode	Random
Packet transfer speed	1-10 mb/s
Traffic model	CBR

5.1 Packet Delivery Fraction

The ratio of the number of data packets received at the destination to the number of those originated at the source.

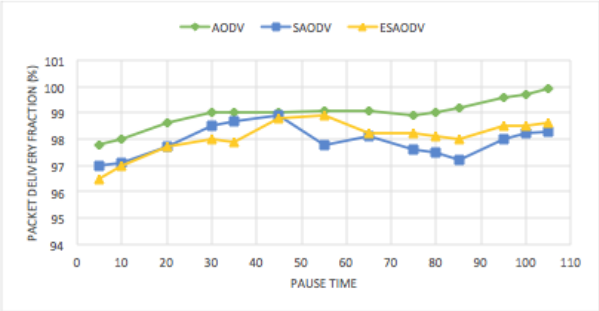


Fig. 5 Packet Delivery Fraction Diagram.

5.2 Average End-to-End Delay

End-to-end delay represents the time that it takes the packets to be received by the destination. Due to the cryptographic techniques in SAODV and ESAODV, average end-to-end delay is more than AODV.

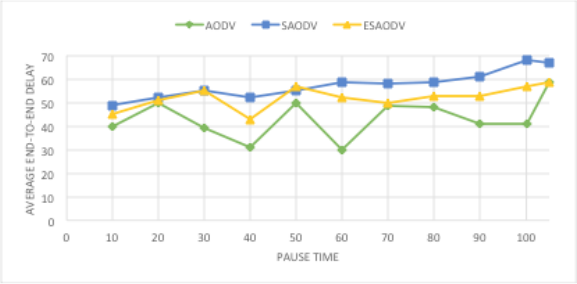


Fig. 6 Average end-to-end delay diagram.

5.3 Number of dropped packet

This parameter represents the number of the dropped packets.

Dropped packet = send packets – receive packets

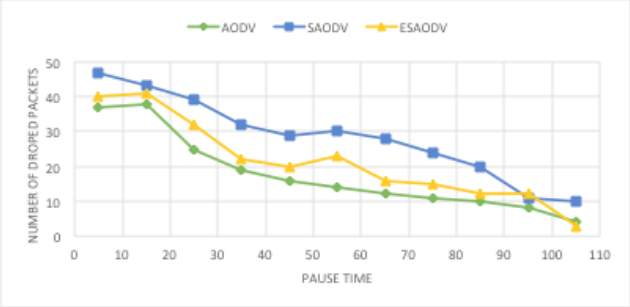


Fig. 7 Number of dropped packet diagram.

5.4 Packet routing overhead

The ratio of the data packet to the number of the routing packets that have been sent.

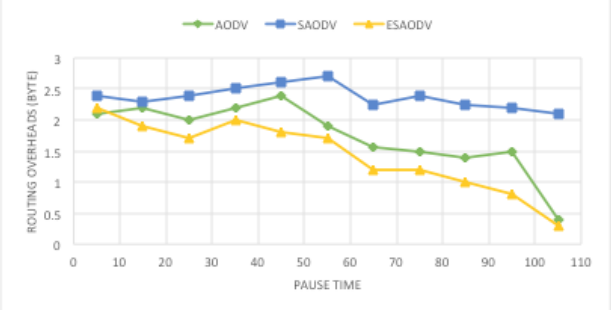


Fig. 8 Packet routing overhead diagram.

Table 1 shows the advantages and disadvantages of some secure routing protocols and the proposed protocol.

Table 2: Advantage and Disadvantage of Secure Routing Protocol and ESAODV

Secure routing protocol	***Prevent from***	***Advantage***	***Disadvantage***
ARAN	Change, Eavesdropping, Impersonation	Easy to implement	Expensive, vulnerable to wormhole
SAR	Change	Dynamic routing, cost	Not always shortest path
SRP	Detection and prevention of impersonated packet	Prevent from eavesdropping	Vulnerable to change and wormhole

SEAD	Change the transferred routing information	Effective use of cpu and energy	Vulnerable to wormhole
ARIADNE	Change, impersonate the routing information	Prevent from wormhole	Abandon malicious node
ESAODV	Change, eavesdropping, black hole	Authenticating the node	Need more storage resources

6. Conclusions

To increment the performance of the network, ESAODV uses the technique in which the intermediate node does not respond to the same RREQ and RREP messages. Also, the fitness function of the genetic algorithm is designed based on minimum delay to find the better path from the source to the destination. Additionally, alternative path prevents from beginning the route discovery process when the source node receives the RERR message. All these mechanisms remarkably decrease the end-to-end delay and the packet routing overhead of the networks.

References

[1] M.Manjunath, D.H. Manjaiah, "Comparative Study of AODV, SAODV, DSDV and AOMDV Routing Protocols in MANET Using NS2," in International Conference on Computing and Intelligence Systems, , March 2015, vol. 4, special issue, pp. 1174-1180.

[2] G. Cerri, A. Ghioni, "Securing AODV: The A-SAODV secure routing prototype," Communications Magazine, IEEE , vol. 46, no. 2, 2008, pp. 120-125.

[3] J. Rajeshwar, Dr. G. Narsimha, "A Comparative Study on Secure Routing Algorithms SAODV and A-SAODV in Mobile Ad-hoc Networks (MANET)- The Enhacements of AODV," International Journal of Computers and Technology, vol. 3, no. 2, 2012, pp. 419-424.

[4] T. R. Andel, "Surveying Security Analysis Techniques in MANET Routing Protocols," IEEE Communications Survey, The Electronic Magazine of Original Peer-Reviewed Survey Articles, vol. 9, no. 4, 2007, pp. 70-85.

[5] J. Neeli, Dr. N.k. Cauvery, "Comparative Study of Secured Routing Protocols in Wireless Ad hoc Networks: A Survey," International Journal of Computer Science and Mobile Computing, vol. 4, no. 2, February 2015, pp. 225-229.

[6] A. k. Mishra, B. D. Sahoo, "A Modified Adaptive-SAODV Prototype for Performance Enhancement in MANET," International Journal of Computer Applications in Enineering, Technology and Sciences (IJ-CA-ETS), vol. 1, no. 2, 2009, pp. 443-447.

[7] A. Banerjee, "Administrator and Trust Based Secure Routing in MANET," in Advances in Mobile Network, Communication and its Applications (MNCAPPS), 2012 International Conference on, 2012.

[8] A. Sharma, M. Sinha, "Influence of crossover and mutation on the behavior of Genetic algorithms in Mobile Ad-hoc Networks," Computing for Sustainable Global Development (INDIACom), International Conference on. IEEE, 2014, pp. 895-899.

[9] J.H. Holland, Adaptation in Natural and Artificial, in The University of Michigan Press, USA, 1975.

[10] D. Suresh Kumar, K. Manikandan, M.A. Saleem Durai, "Secure on-demand Routing Protocol for MANET Using Genetic Algorithm," International Journal of Computer Applications, vol. 19, no. 8, April 2011, pp. 29-35.

[11] E. Baburaj, V. Vasudevan, "An Intelligent Multicast Ad-hoc On demand Distance Vector Protocol for MANETs," Journal of Networks, vol. 3, no. 6, June 2008, pp. 62-69.

[12] P. Gaur, "An Efficient Routing Implementation Using Genetic Algorithm," International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 2, no. 7, 2013, pp. 250-257.

[13] P. Singh, G. Singh, "Security issues and link expiration in secure routing protocols in MANET: a review," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 7,July 2014, pp. 7559-7565.

Atieh Moghaddam is completed the Bachelor of Science in software engineering in 2010. She completed the Master of Science in information security from Tehran University in 2013. She is a member of network Security Company. Her area of interest spans Genetic Algorithm, security issues in mobile ad-hoc networks and cryptography.

Ali Payandeh received B.Sc. and M.Sc. degrees in electrical engineering From Tarbiat Modarres University, Iran, in 1991 and 1994, respectively, And the Ph.D. degree in electrical engineering from K. N. Toosi University of Technology, Iran, in 2006. From 1991 to 1995, he was a faculty member in the Department of Electrical engineering at Malek-e-Ashtar University of technology, Iran. Since 1996, he has been a Director of Research at the Applied Science Research Association (ASRA), Iran, where he has involved in research for secure satellite communications. His research interests include information theory, coding theory, secure communications and satellite communications.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

MAS-based auction for channel selection in mobile cognitive radio networks

Emna Trigui, Moez Esseghir, Leila Merghem-Boulahia

ICD/ERA, CNRS UMR ICD 6281, University of Technology of Troyes,
12, rue Marie Curie, 10010 Troyes Cedex, France
{emna.trigui, moez.esseghir, leila.merghem_boulahia} @utt.fr

Abstract

Cognitive radio network is a concept of wireless communication for mobile devices that offers the possibility to exploit the unused spectrum resources opportunistically. These networks bring out the need for new solutions that mitigate the spectrum management issue. However, existing works do not focus on devices mobility whereas serious problems arise when users are mobile specifically about their provided quality of services. In this work, we study spectrum sharing and spectrum handoff for mobile secondary users (SUs) and we propose a novel approach that can be executed by a mobile SU when traveling through wireless networks. The proposed solution is inspired from multi-agent system auctions and integrates a learning module which accelerates SUs' spectrum bands allocation. One of the main contributions of this paper is the realistic implementation of the learning based auction and the interesting results obtained through a network discrete event simulator. Results prove that our proposal enhances spectrum utilization and guarantees users satisfaction.

Keywords**:** *Auction, Spectrum Access, Mobility, Cognitive radio, resources management, learning*

1. Introduction

In the last decade, cognitive radio [1, 2] technology has received a tremendous attention thanks to its opportunistic spectrum access abilities and its reconfiguration capabilities. A cognitive radio network is a set of wireless devices that tries to access the spectrum resources opportunistically. These devices are known as secondary or unlicensed users (SUs). Licensed devices, known as primary users (PUs), will share license spectrum with SUs.

Spectrum management task is an important challenge in a CR network as it includes the four main functionalities of a cognitive radio (CR) device: (1) spectrum sensing to detect spectrum holes; (2) spectrum decision to select the most appropriate frequency band; (3) spectrum sharing; (4) and finally spectrum handoff to switch channel whether it is necessary.

Node's mobility magnifies the spectrum management problem since user's handover can badly affect the provided quality of service (QoS). Consequently, the need of complementary researches in CR spectrum handoff is extremely important.

This work aims to provide a seamless spectrum handoff while ensuring efficient spectrum allocation for mobile CR users.
The major contributions of this paper are as follows.

1) We propose a multi-agent system based auction algorithm for both spectrum sharing and handoff decisions.
2) We derive a realistic implementation that can be easily deployed on PUs and SUs.
3) We enhance the system performances using a straightforward learning module.

Broadly, existing works use analytical approaches and game theory solutions which produce theoretical results. However, the need for more easily deployable, distributed, and scalable solutions is highly relevant. For this reason, we rely on multi-agent system (MAS).

The remainder of the paper is organized as follows. Section II describes recent works on auction based spectrum management in CR networks. Section III details our proposed approach. We present the context and the MAS-based auction we propose for handoff and spectrum access. We depict both PU's and SU's behaviors with the optional learning. Section IV gives the extensive simulation results and section V concludes the paper.

2. Related word

Extensive literature is available on the study of dynamic spectrum management [3, 4] in cognitive radio networks using various mechanisms. Among the different mechanisms proposed to address spectrum allocation, an effective technique has been the use of auctions [5]. There is substantial agreement among economists that auctions are the best way to assign scarce resources [6].

Furthermore, spectrum trading via auctions allows a more dynamic, competitive and efficient communications market than is possible under the traditional systems implemented so far, mainly because spectrum users and wireless service providers have better knowledge than regulators about their spectrum requirements and valuations. For further details, we can refer to the survey of auction mechanisms designed for dynamic spectrum allocation in [7] and the tutorial paper in [8], which discuss the use of auctions for dynamic spectrum allocation in CR networks. Details on other schemes that have been proposed to address the problem of dynamic spectrum management are available in the surveys [9] and [10].
Among the limitations of existing work using auction theory for spectrum allocation, we quote the extensive use of analytical approaches and game theory solutions which

produce theoretical results while we need for more deployable solutions.

For example, in [11] the auction occurs between one PU and multiple SUs sharing the same spectrum in a CR network. Each SU makes a bid for the amount of spectrum it requires and the PU assigns the spectrum band to the SU that do not damage its quality of service (QoS). The objective of this study was to find the Nash Equilibrium (NE) state. In [12], authors formulate the problem as a non cooperative auction game and study the structure of the resulting NE by solving a non-continuous two dimensional optimization problem. Each SU updates its strategy based on local information to converge to the NE. This study can theoretically serve as a decision and control routines for the SUs to exploit the underutilized spectrum resource.

A further limitation in existing auction based spectrum management researches consists in focusing simply on a MAC layer solution ignoring the rest of layers.

In [13], for instance, a Q-learning based bidding algorithm for spectrum auction is proposed, which enables SUs to bid for available frequency bands automatically. This study presents a bidding algorithm for SUs in each time slot. Authors study buffering and channels occupation and they are not interested in the pricing issue for spectrum bands allocation.

Authors in [14] propose a cognitive MAC protocol for CR networks on the basis of the combinatorial auction principle. Moreover, both of the two designs proposed in [15] and [16] are based on analytical analysis to solve channel access at the level of low layers.

In this paper, we implement the First Price Sealed Bid auction at the application level with a real billing system to provide an easily deployable and scalable solution. Furthermore, we have introduced an effective solution for channel selection and spectrum access by considering users' mobility. We have integrated a learning module to enhance system performances.

3. Novel auction based protocol

In this section, we briefly describe the scenario we use and we present our proposed approach for spectrum access and handoff in mobile CR networks.

We propose a solution for spectrum management at the application layer where we integrate the selection and learning modules. We are referred to the IEEE 802.11 standard for the physical and MAC layers and we consider the IP protocol at the network layer. Fig. 1 shows our model's architecture from a stack layers point of view. .

We keep our protocol general so that it can be applied in any current or future system equipped with CR technology as IEEE 802.11af or IEEE 802.22 standards, which have advocated using white spaces left by the termination of analog TV to provide wireless broadband internet access. A device intended to use these available channels is called a "white-spaces device" (WSD). In our model, SUs have the abilities to be WSDs. The spectrum is located in the VHF/UHF bands (470-806 MHz) and has the characteristics that make it highly desirable for wireless communications.

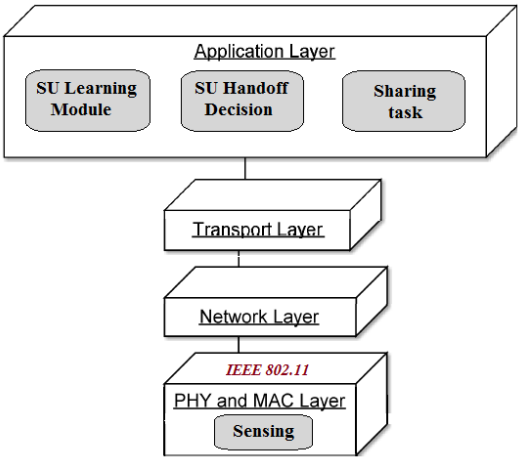


Fig.1. Model's Architecture

For this work, we consider ad-hoc network with a set of primary and secondary users. SUs are mobile nodes and PUs are fixed ones. Each node is operating in a frequency band and each PU can have unused frequencies (sub-bands). With cognitive radio technology, nodes become able to switch from one frequency to another. When an SU is moving from one zone to another one, available resources may change and the CR node will be able to use another available spectrum band.

The challenge in the previous scenario consists in allowing SUs to select the target channel promptly and to move from one zone to another one seamlessly without causing service interruption.

Spectrum handoff and allocation processes will be modelled through an auction between PUs and SUs existing in the same zone. Each PU having free bands starts an auction and is considered to be the auctioneer. On the other side, SUs are the bidders and try to submit their offer until a potential win. Besides, the proposed auction is improved with a learning module to enhance the system efficiency and users' band attribution.

The price for spectrum band access is determined by CR users (i.e., bidders). The multiunit sealed-bid auction as the first price sealed-bid auction (FPSB) is very suitable to execute in a determinable time with an acceptable signaling effort in comparison to the sequential auction such as the English one. In addition, the FPSB allow assigning spectrum holes to CR

users faster than the traditional English auction as the FPSB is a single round auction however the English one is a multiple round auction [17]. For these reasons, we use the FPSB in our proposed solution.

Accordingly, all bidders (SUs) simultaneously submit their sealed bids. The highest bidder wins and the corresponding SU pays its submitted bid. Fig.2 illustrates the considered FPSB auction between PU and SUs agents.

First, each PU initiates an auction when some of its licensed bands is released. It forwards a START_AUCTION message to neighbouring CR users. An SU needs to access spectrum in two cases: (1) When coming close to a new zone where the radio resources change; (2) When its attributed spectrum use duration expires.

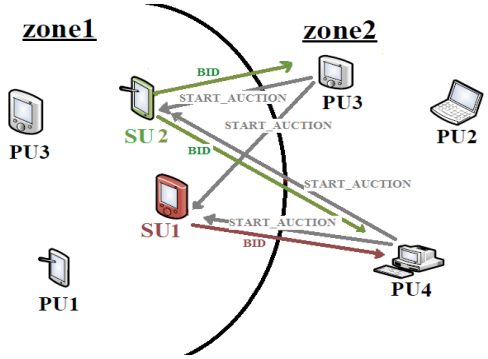


Fig.2. Auction modeling

The START_AUCTION message sent by the PU contains its licensed frequencies (*Freq(PU)*) as well as the amount of free spectrum sub-bands that can be allocated ($S_{free}(PU)$). Whenever this amount of available sub-bands covers an SU's needs in terms of spectrum resources ($S_{Needed}(SU)$), this SU participates in the initiated auction and sends a BID message containing its offer in terms of unit price per second (*PPS*). The other SUs who need more spectrum sub-bands do not participate and wait for another auction.

The PU auctioneer selects among received bids the one that presents the highest price per second and sends WINNER message to the corresponding SU (SU_w:*SU winner)* in order to start sharing bands. The PU assigns its proposed use duration (*D(PU)*) for a price (P_{paid}). Each PU has its own fixed *D(PU)*. Use duration can be different for each PU. P_{paid} is calculated as a function of the price per second proposed by the SU winner ($PPS(SU_W)$). The PU waits for a positive acknowledgment (*ACK(OK)*) from the SU_W to start sharing.

The SU shares bands with the PU that answers the first. If it receives another WINNER message later from other PUs while it is already sharing a PU's bands, the SU withdraws and sends a negative acknowledgment (*ACK(NO)*) to precise that it has already won an auction. In this case, the negatively notified PU restarts the auction process to choose another available winner.

In case of positive ACK, the PU shares the required sub-bands with the SU winner and restarts another auction if it still disposes of free sub-bands. Note that the PU's own spectrum bands utilization varies over time.

In the following, we present in details the sequence diagram, the PU's behavior and SU's algorithm. We depict the proposed algorithms dealing with their different steps.

3.1 Sequence Diagram

Fig. 3 describes the sequence diagram between a PU (the auctioneer) and an SU (a bidder). The PU forwards a **START_AUCTION** message and waits for SUs bids. This call for auction contains the licensed frequencies and the amount of free sub-bands of the PU. Interested SU responds with a **BID** message containing its PPS_{bid}. The PU sends **WINNER** message with the proposed duration (*D(PU)*) to the SU offering the highest PPS. If the selected SU has not won another auction (no sharing band), it responds with positive ACK and a sharing band starts. Otherwise, it sends a negative ACK. In this case, the PU has to restart another auction.

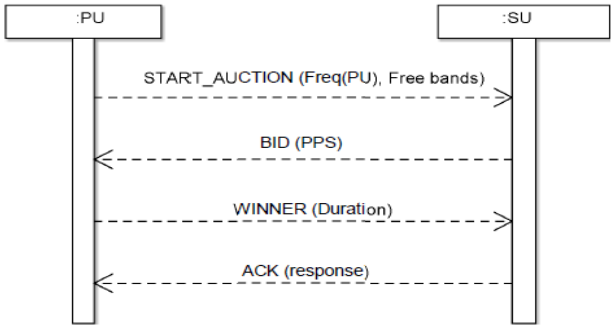


Fig.3. Sequence diagram

3.2 PU behavior algorithm

The PU behavior is described in the algorithm A.

Whenever a PU starts an auction, it waits for a given time to receive SUs' bids. This waiting time is noted Γ.

Each PU has a Price Per Second Reserve $PPS_{reserve}$. The PU cannot accept an offer (*PPS*) lower than its $PPS_{reserve}$. This later is given by equation (1).

$$PPS_{reserve}(PU_i) = \frac{PR_i(PU_i)}{DR_i(PU_i)} \quad (1)$$

$$\text{Where } \begin{cases} PR_i(PU_i) \in [PR_{min}(PU), PR_{max}(PU)] \\ DR_i(PU_i) \in [DR_{min}(PU), DR_{max}(PU)] \end{cases}$$

$PR_{min}(PU)$ and $PR_{max}(PU)$ represents the minimum and the maximum prices that can propose a PU for its bands allocation. They are fixed values and are the same for all PUs.

Algorithm A: PU behavior

```
BEGIN
Sharing == false //The PU is sharing its bands
FreeBands=true //The PU has free bands
Repeat
  If (FreeBands==true)
  Then // The PU forwards START_AUCTION message
     FWD (START_AUCTION_MSG)
    //The PU waits for a given time (Γ) to receive bids
     While (Γ not expires)
        If (receive bid)
        Then
            Insert the bid in Bids_Vec
        End If
     End While
     If ( Γ expires)
     Then  //The PU selects the SU winner
        For i = 1 … Nb // Nb number of bids; Nb =Bids_Vec.Size()
           If ( PPSi < PPSreserve ) // PPSi :the bid number i
           Then
              Elimination of PPSi bid
           End If
        End For
        SUWinner ← SU that proposes the highest PPS
        Ppaid ← D(PU) * PPS(SUW)* SNeeded(SUW)
        //The PU send WINNER message with the attributed duration
        SEND (WINNER_MSG) to SUW
     End If
     If (SU's ACK == OK)
     Then
        Sharing ← true // sharing band
      //Test if SU's band needed (SNeeded) is less than PU's free band (Sfree)
        If (SNeeded(SU)< Sfree(PU))
        Then
           FreeBands ← true
        Else
           FreeBands ← false
        End If
     Else  //SU is sharing another band
        Sharing ←false
     End If
  End If
Until (Sharing == true and FreeBands == false)
END
```

$DR_{min}(PU)$ and $DR_{max}(PU)$ represents the minimum and the maximum use durations that can attributes a PU for bands allocation. Likely, these values are fixed and the same for all PUs.

From received bids in Γ time, the PU eliminates the bids where the *PPS* is lower than its $PPS_{reserve}$. Then, it chooses the SU that proposes the highest *PPS*. The PU sends a WINNER message to this selected SU (SU_W) for a spectrum sharing with the price P_{paid} given by the following equation (2) and for the use duration *D(PU)* initially proposed by the PU.

$$P_{paid} = D(PU) * PPS(SU_W) * S_{Needed}(SU_w) \quad (2)$$

Where $PPS(SU_W)$ is the unit Price Per Second of the SU that wins the auction and $S_{Needed}(SU_W)$ is the amount of spectrum bands needed by SU_W .

If the PU receives the same offer more than once from two or more different SUs (i.e. same *PPS*), the PU chooses one of them randomly. The following sub-section details the SU's algorithm.

3.3 SU behavior algorithm

Algorithm B: SU behavior

```
BEGIN
HO_Var = true //The SU is switching network
InShare = true /The SU is sharing PU's bands
ηA(PUi) = 0 // Number of received Auction from the same PUi

// If the SU is changing zone, it has to search for another free band
If (HO_Var == true)
Then
   InShare ←false
End If
If ( (InShare == false) and reception of START_AUCTION (Freq(PUi), Sfree(PUi)))
Then  // The SU verifies if the auction call propose sufficient free bands
   If ( Sfree(PUi) ≥ SNeeded(SU))
   Then // Learning Module
      // ηA is the number of auction calls received from the same PUi
      ηA(PUi) ← ηA(PUi) + 1
      If ( ηA(PUi) > 1)
      Then
         PPSbid ← PPSinitial(SU) + ηA(PUi) * ψ
         If ( PPSbid >PPSmax)
         Then
            PPSbid ← PPSmax
         End If
      Else
         PPSbid ← PPSinitial(SU)
      End If
      SEND_BID (PPSbid)
   End If
End If
If (received WINNER_MSG(duration(PU)))
Then
   If (InShare ==false) // The SU reply by a positive acknowledgment
      SEND(ACK(OK))
      ηA(PUi) ← 0
      Ppaid ← duration(PU) * PPS(SU)* SNeeded(SU)
      Freq(SU) ← Freq (PU) //Spectrum handoff
      InSahre ← true
   Else //The SU reply by a negative acknowledgment
      SEND(ACK(NO))
End If
If ( duration expires)
Then
      InShare ← false
End If
END
```

Once an SU comes close to a new zone, it waits for incoming auction calls. When it receives a START AUCTION message, the SU verifies if the PU initiating this auction offers sufficient sub-bands as it demands. The SU participates only to auctions where the PU's available bands ($S_{free}(PU)$) cover its requirements ($S_{Needed}(SU)$). The SU takes part in all auctions that satisfy its needs until it is selected as winner of an auction. In each involvement, the SU proposes a bid in terms of *PPS*.

Each SU_i has a price $P_i(SU_i)$ to provide for spectrum allocation, it has also a favorite duration $D_i(SU_i)$. Hence, it has an initial price per second noted $PPS_{initial}(SU_i)$, given by equation (3).

$$PPS_{initial}(SU_i) = \frac{P_i(SU_i)}{D_i(SU_i)} \quad (3)$$

$$\text{Where } \begin{cases} P_i(SU_i) \in [P_{min}(SU), P_{max}(SU)] \\ D_i(SU_i) \in [D_{min}(SU), D_{max}(SU)] \end{cases}$$

$P_{min}(SU)$ and $P_{max}(SU)$ represents the minimum and the maximum prices that can bid an SU for bands allocation. $D_{min}(SU)$ and $D_{max}(SU)$ are the minimum and the maximum use durations that can demand an SU for bands allocation. All SUs have the same prices and use durations' bounds.

Firstly, we have implemented our proposed auction based spectrum management protocol considering that each SU sends its PPS_{intial} (as a bid). This case is referred as without learning:

$$PPS_{bid}(SU_i) = PPS_{initial}(SU_i)$$

Then, we have integrated a learning module in the SU's behavior to increase each SU chance to win auctions and access the spectrum more quickly. The following sub-section describes our used learning module, which is for this study, straightforward.

Learning process

We define $PPS_{max}(SU)$ as the maximum price per second that can propose an SU for spectrum allocation. The PPS_{max} is given by equation (4):

$$PPS_{max}(SU) = \frac{P_{max}(SU)}{D_{min}(SU)} \quad (4)$$

$P_{max}(SU)$ and $D_{min}(SU)$ are previously defined as follows

$$\begin{cases} P_{max}(SU) = Max(P_i(SU_i)) \ \forall i \\ D_{min}(SU) = Min(D_i(SU_i)) \ \forall i \end{cases}$$

The key idea behind our Learning module is to increase the SU's bid (PPS_{bid}) whenever the SU receives an auction re-call from the same PU provided that the new PPS_{bid} does not exceed PPS_{max}.This increase will be modelled by the learning parameter noted ψ.

The PPS_{bid} will depend on the number of auction calls received from the same PU_i. This number is noted $\eta_A(PU_i)$. $PPS_{bid}(SU)$ is calculated by equation (5).

$$\begin{cases} PPS_{bid}(SU) = PPS_{initial}(SU) + \eta_A(PU_i) * \psi \\ PPS_{bid}(SU) \le PPS_{max} \end{cases} \quad (5)$$

The SU resets its PPS_{bid} to its $PPS_{initial}$ after each successful band sharing.

We assume that the SU is changing its environmental parameter (HO_Var←true in the algorithm B) automatically when it comes close to a new zone. The SU anticipates changing zone when its average Received Signal Strength (RSS) becomes lower than the RSS limit, which ensures a good QoS. The SU behavior is detailed in the algorithm B.

To evaluate the performances of our proposed protocol, extensive tests are conducted. In the following section we analyse the simulation results.

4. Result

We perform our tests under OMNETPP simulator [18], which is a discrete event simulation network tool.

We consider the specific case where SUs move from an initial zone to a second one. We randomly deploy PUs over these two zones and SUs arrive following a Poisson distribution with parameter λ set to 5. We suppose that SUs are continuously requiring spectrum access. Spectrum is divided in equal bands of 4 MHz bandwidth. Each band is sub-divided into 4 equal sub-bands of 1 MHz. We assume that a PU can own 0 to 4 free sub-bands. The number of simulation runs is set to 10 and the results are averaged to plot graphs. In all our simulations, a 95% confidence interval is computed for each average value represented in the curves. These intervals are plotted as error bars. The rest of the simulation parameters are given in table1.

TABLE 1
SIMULATION PARAMETERS

Parameters	Values
SU distribution (λ)	5
PU number (nb_{PUs})	100
SU number (nb_{SUs})	{100, 110, 120, 130,140,150,160}
Bid Waiting Time (Γ)	0.6 s
Learning parameter (ψ)	{0.1, 0.2, 0.3, 0.4, 0.5}
$P_i(SU) \epsilon [P_{min}(SU), P_{max}(SU)]$	[30, 50] (unit price)
$D_i(SU_i) \epsilon [D_{min}(SU), D_{max}(SU)]$	[45, 120] (unit time)
$PR_i(PU_i) \epsilon [PR_{min}(PU), PR_{max}(PU)]$	[35, 55] (unit price)
$DR_i(PU_i) \epsilon [DR_{min}(PU), DR_{max}(PU)]$	[60, 240] (unit time)
Size of spectrum band	4 MHz
Size of spectrum sub-band	1 MHz
SU speed	10mps
SU Mobility type	Linear
Simulation time	600 s
Simulation runs number	10

First, we have evaluated the implementation of our auction based approach without learning. We present the spectrum utilization and the handoff delay, then we study the impact of the $PPS_{reserve}(PU)$ on the handoff blocking rate, on users' utility and on Handoff delay.

Next, we evaluate the impact of learning contribution on the performance of the spectrum management protocol. For that, we have compared obtained results when we integrate the learning module with the case *without learning*. Besides, we study the impact of the learning parameter (ψ) on the system performances.

4.1 Auction implementation results

In this subsection, we present results of the basic auction protocol (without learning). We introduce first the spectrum utilization over time. Then, we expose Handoff delay, the average blocking rate and users' utility as a function of SUs number.

Spectrum utilization

The spectrum utilization rate is equal to the amount of spectrum bands utilized by all PUs and all SUs present in the same zone divided by the total amount of existing bands. Fig.4 shows the average rate of spectrum use within 600s for a total of 130 SUs and 100 PUs.

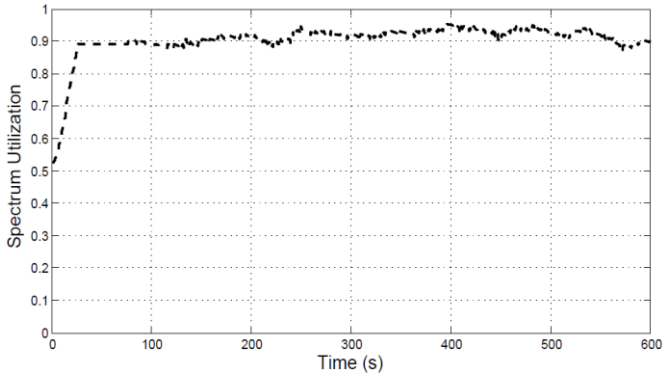


Fig.4. Spectrum utilization rate over time

We observe from Fig.4 that the spectrum utilization rate can achieve up to 94% of the whole available spectrum and then reaches a steady state in a transient time until the end of the simulation. This proves clearly that our protocol improves significantly the spectrum use.

In the next sections, we will study the impact of the number of SUs on the handoff delay and the blocking rate.

Handoff Delay

The Handoff delay (D_{HO}) in these analyses is calculated as the average waiting time between two successive spectrum accesses. The Handoff delay is given by equation (6).

$$D_{HO} = \frac{1}{nb_{SUs}} * \sum_{nb_{SUs}} [\frac{1}{N_{all}} * \sum_{i=1}^{N_{all}} (T_A(B_{i+1}) - (T_E(B_i)] \quad (6)$$

Where $T_A(B(_{i+1}))$ is the time allocation of a band (i+1) and $T_E(B_i)$ is the end time of the i^{th} band used by SU. N_{all} is the total number of spectrum allocations for SUs and nb_{SUs} is the number of SUs present in the system.

The bar chart in Fig. 5 presents the handoff delay as a function of SUs number compared to the average spectrum use duration that can an SU obtain. Fig.5 shows also the rate of SUs that have successfully access the spectrum.

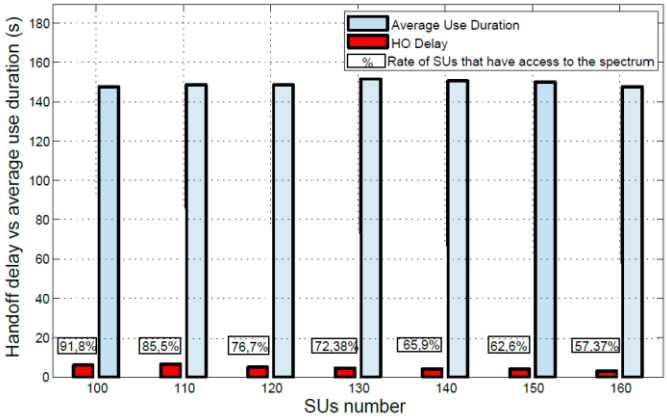


Fig.5. Handoff delay versus the average use duration

The most interesting result is that the handoff delay is extremely low compared to the average spectrum use duration. This result proves that the proposed auction based approach ensures low interruption time and guarantees service continuity.

Besides, we observe that the handoff delay decreases slightly when the number of SUs increases. This is explained by the fact that the percentage of SUs successfully accessing the spectrum decreases. For example, with 100 SUs there are an average of 91.8 SUs that have successfully access the spectrum resources. Consequently, the handoff delay presented is relative to 91.8% SUs.

Blocking rate

To evaluate the smooth functioning of the proposed system, we measure the percentage of SUs that have failed to use the spectrum, i.e. SUs that lost all tripped auctions. This percentage is noted blocking rate and is plotted in Fig. 6 as a function of SUs number.

Furthermore, we study the impact of PU's $PPS_{reserve}$ on the blocking rate and then on users' utility. We consider three cases of $PPS_{reserve}$: a random case and two boundaries values of $PPS_{reserve}$, respectively ($Min(PPS_{reserve})$) and $Max(PPS_{reserve})$. Note that $Min(PPS_{reserve})$ and $Max(PPS_{reserve})$ are as follows:

$$Min(PPS_{reserve}(PU)) = \frac{PR_{\min}(PU)}{DR_{\max}(PU)}$$

$$Max(PPS_{reserve}(PU)) = \frac{PR_{\max}(PU)}{DR_{\min}(PU)}$$

Fig. 6 shows that the blocking rate increases considerably when the $PPS_{reserve}$ is equal to Max($PPS_{reserve}$). This result is expected since the $PPS_{reserve}$ in this case is generally higher than the average PPS proposed by the SUs. Consequently, PUs will eliminate most received bids and few SUs access successfully to the spectrum. However, we clearly observe that the blocking rate is notably lower when the $PPS_{reserve}$ is the minimum. It is important to note that when the $PPS_{reserve}$ is random value, i.e. the general case, we obtain low blocking rate near to the minimum case. This proves that our approach ensures a significant exploitation of the spectrum resources and can satisfy the needs of most SUs.

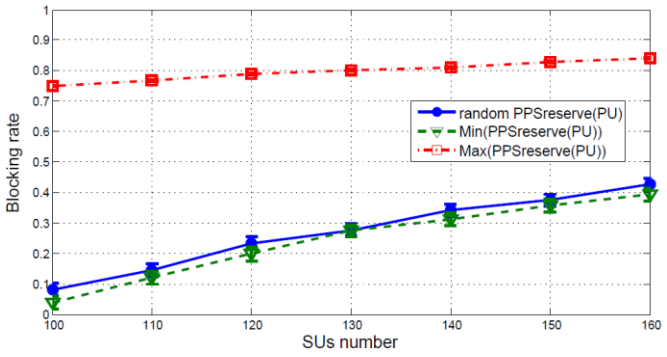


Fig.6. Blocking rate

The blocking rate increases when the number of SUs rises, which is expected since the available resources are unchanged and first coming SUs will be the first served. Thus, the probability to receive a call for auction that presents sufficient free bands becomes too low when the number of SUs increases.

Users Utility

User's utility is a very important metric to evaluate the satisfaction of the network's users. Therefore, we have measured CR users' utility as well as PUs' utility. We have also studied the impact of the $PPS_{reserve}$ on both measures.

SUs' utility

In this scenario, the SUs' utility can be defined as the SUs' benefit from PUs offers. In other words, an SU wants to have more spectrum use duration with a minimum price. The utility of the i[th] SU noted $U(SU_i)$ is given by equation (7).

$$U(SU_i) = \frac{1}{N_{all}} \sum_{N_{all}} \left(\frac{PR_{\min}(PU)}{P_{Unit\,paid}} * \frac{D_{attributed}}{DR_{\max}(PU)} \right) \quad (7)$$

Where N_{all} is the total number of successful spectrum allocation of the SU_i in the simulation. $P_{Unit\,paid}$ is the unit price paid (for a sub-band allocation) and $D_{attributed}$ is the attributed duration for the spectrum access. Recall that $PR_{min}(PU)$ and $DR_{max}(PU)$ represents the minimum price and the maximum use duration for PU's allocated spectrum bands, respectively.

The SU's utility can be presented otherwise, as a function of the SU's proposed PPS. $U(SU_i)$ can be given by the following:

$$U(SU_i) = \frac{1}{N_{all}} \sum_{N_{all}} \left(\frac{Min(PPS_{reserve}(PU))}{PPS_{bid}(SU_i)} \right) \quad (8)$$

Fig.7 shows the average SUs' utility in the previous three cases of $PPS_{reserve}$. We observe that our proposed auction protocol ensures a good SUs' utility when we consider flexible PUs (random $PPS_{reserve}$) very nearly to the case of non-strict PUs (Min ($PPS_{reserve}$)) and is largely better than obtained SU's utility with strict PUs (Max ($PPS_{reserve}$)).

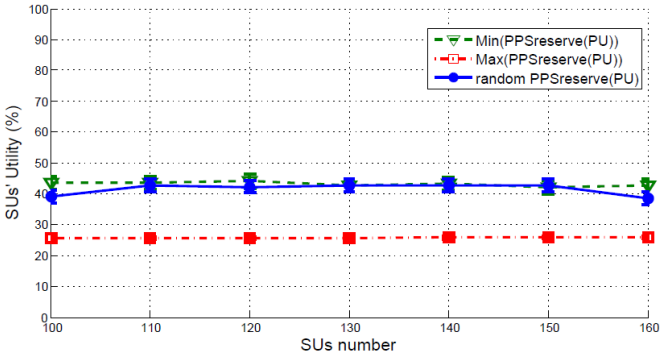


Fig.7. Impact of $PPS_{reserve}$ and SUs number on the SUs' utility

In the next subsection, we present the impact of the PU's flexibility (i.e. boundaries of $PPS_{reserve}$) in its average utility.

PUs' utility

We define the PUs' utility as the PUs' profit from SUs' bids. The utility of the i[th] PU noted $U(PU_i)$ is given by the equation (9).

$$U(PU_i) = \frac{P_{Unit\,paid}}{P_{\max}(SU)} * \frac{D_{\min}(SU)}{D_{attributed}} \quad (9)$$

Where $P_{max}(SU)$ and $D_{min}(SU)$ represents the maximum price and the minimum favorite use duration that can propose an SU for spectrum allocation respectively.

PU's utility can be presented otherwise, as a function of the SU winner's proposed bid (PPS_{bid}) and inversely proportional to the SUs' PPS_{max} as given by equation (10).

$$U(PU_i) = \frac{PPS_{bid}(SU_W)}{PPS_{\max}(SU)} \quad (10)$$

Fig. 8 draws the PUs' utility as a function of SUs number. It shows that PUs' utility is more important when the PU is very strict (Max ($PPS_{reserve}$)), which is obvious since the PU accept only bids offering very high PPS. Our proposal with random $PPS_{reserve}$ ensures important PUs' satisfaction that reaches 70%.

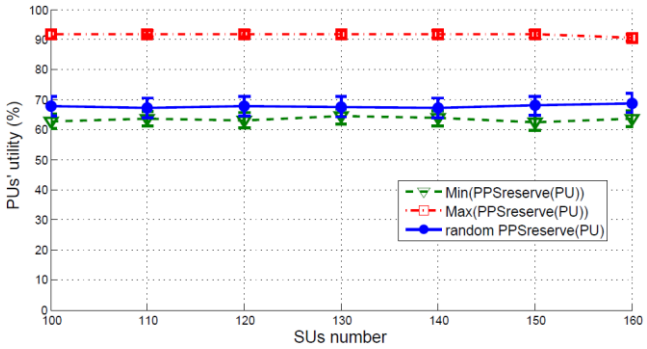


Fig.8. Impact of $PPS_{reserve}$ and SUs number on the PUs' utility

Since we showed above, the implemented auction protocol provides high spectrum utilization, low blocking rate and ensures users' satisfaction. $PPS_{reserve}$ study proves that the proposed bids (SUs' PPS) should not be far from the $PPS_{reserve}$ to have efficient system. This condition is generally satisfied since bidders in auction market know approximately the price range of the proposed product.

Hence, it is widely interesting to involve the learning process into CR devices. Whenever its bid is rejected, the SU tries to increase it so as to reach the $PPS_{reserve}$.

The remainder of conducted simulations is devoted to study the impact of the integrated learning module in our proposed spectrum management based auction protocol.

4.2 Learning based auction for spectrum management results

In this section, we compare the two alternatives of the proposed auction based protocol, one with learning module and the second without learning. We study the impact of the SUs number as well as the learning parameter ψ on some important metrics such as handoff delay and users satisfaction.

Average number of attempts before spectrum access

One of the major objectives when introducing the learning process is to accelerate SUs' spectrum access. Thus, we measure the average number of SUs' attempts (failed bids) before spectrum access (i.e. before auction win).

First, we assume the learning parameter ψ is equal to 0.1 in equation (5) and study the average number of attempts before spectrum access as a function of SUs number. Results are shown in Fig. 9.

Fig. 9 clears that the learning module decreases extremely the average number of SUs' attempts before spectrum access. This number is reduced from more than 120 to 25 attempts.

This important result proves that our proposal of using learning based auction for spectrum management enhances largely the bidding efficiency and the access opportunity.

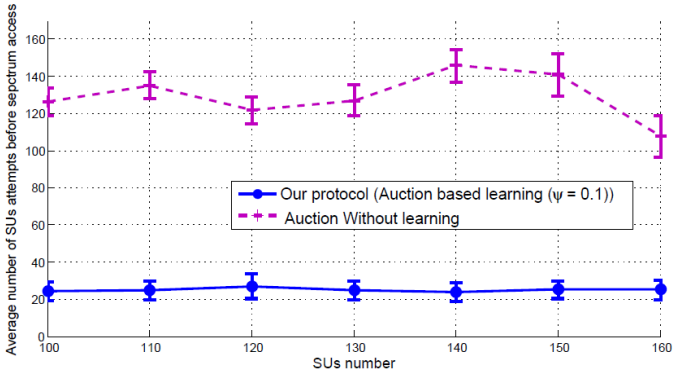


Fig. 9 Average number of SUs' attempts before spectrum access as a function of SUs number

The learning process is modeled through the learning parameter ψ. Consequently, we study the impact of ψ on the average number of attempts before spectrum access as shown in Fig.10. We fix the SUs number to 120 and we vary the ψ parameter between 0.1 and 0.5.

Fig.10 proves that increasing the learning parameter allows to further reduce the average number of attempts before spectrum access. This is explained by the fact that the SUs' bids reach the PUs' $PPS_{reserve}$ more quickly.

Besides, we present in Fig.10 the auction based protocol when considering that all SUs send the same bids, equal to PPS_{max}. When the learning parameter increases, the average number of attempts obtained with our protocol is approaching obtained results when SU's PPS_{bid} is equal to PPS_{max}.

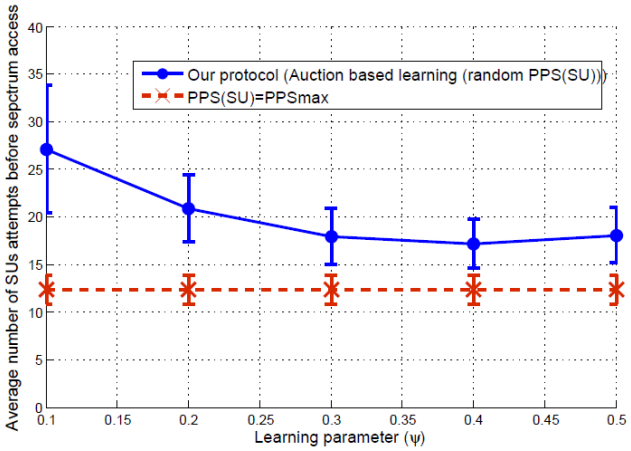


Fig. 10 ψ impact on the average number of SUs' attempts before spectrum access

In the next subsections, we study the impact of the learning process on blocking rate, handoff delay and users' utility.

Average blocking rate

Another objective of the learning module is to make more SUs able to access the spectrum. To confirm this property, we have measured the blocking rate that reflects the percentage of SUs that have failed to access the spectrum. Fig. 11 presents the comparison results between the learning based approach and without learning alternative.

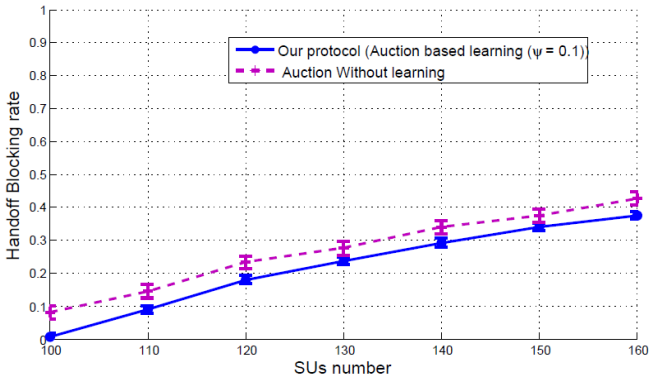


Fig. 11 Impact of learning module on the blocking rate

Fig.11 shows that the learning module decreases the blocking rate, which proves that the learning based auction proposal improves the number of SUs accessing the spectrum resources. As previously explained (section IV.A.3) the blocking rate increases when the SUs' number rises. This is due to the scarcity of free spectrum bands when having a large number of SUs.

Users' Utility

Learning module can impact on users' utility as the SUs will send higher bids and PUs will receive more interesting offers. Thus, a priori, the PU will be the beneficial from the learning process in terms of reward. Effectively, as confirmed by results in Fig. 12, PUs' utility is enhanced inversely to the SUs' utility.

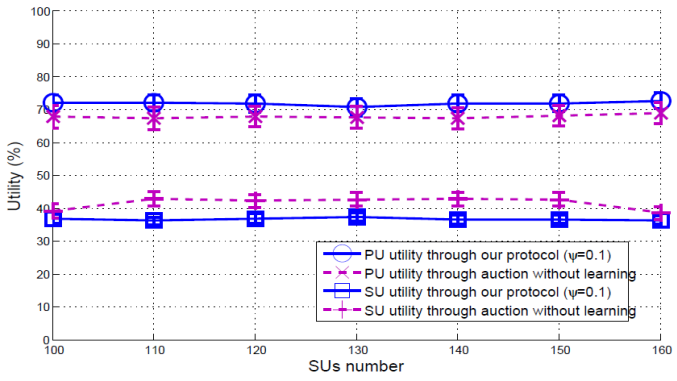


Fig. 12 Impact of learning module on users' utility

To reveal the learning parameter influence on users' utility, we varied ψ while fixing the SUs number to 120. Fig.13 presents the comparison between the obtained results with random $PPS_{reserve}$ and the case where all bids are maximized ($PPS_{bids}=PPS_{max}$).

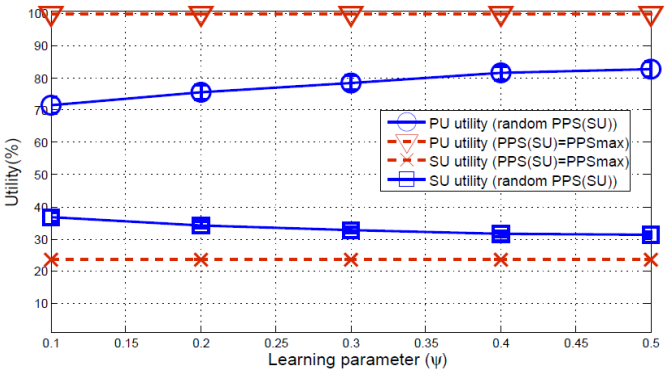


Fig. 13 Impact of ψ parameter on users' utility

Fig.13 shows that the PUs' utility increases when ψ rises. Contrary to SUs' utility that decreases when ψ increases and tends towards the lowest utility that can be obtained in the case of sending the PPS_{max}.

Handoff Delay

Another important metric that must be studied when considering learning module is the handoff delay. We present in Fig.14 the corresponding results. We observe through Fig.14 that the handoff delay increases slightly with the learning process. This is obvious and expected because the number of SUs successfully access the spectrum increases. This involves more spectrum bands occupancy and more time for channels release. Consequently, SUs will need additional time to perform handoff and to access the spectrum.

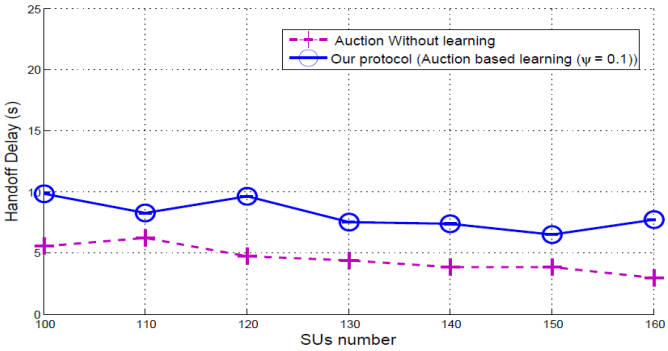


Fig. 14 Impact of learning module on the handoff delay

To summarize, the implemented learning based auction protocol for spectrum access and channel selection ensures an important spectrum exploitation and low handoff delay compared to the average use duration. Our proposal guarantees likewise good utility for both primary and secondary users.

5. Conclusions

In this paper, we designed a novel multi-agent based auction system for spectrum allocation and channel selection. We have improved our approach through a straightforward learning module. Besides, our proposal integrate a real billing and pricing system that can be easily deployed in actual and future wireless networks.

Simulation results prove that our proposal provides high spectrum utilization, low blocking rate and ensures users' satisfaction. Furthermore, we showed that the learning module improves mobile cognitive radio users' behavior in terms of bidding efficiency and access opportunity. In addition, it enhances the primary users' utility and reduces the overall blocking rate.

As future work, we intend to study different learning strategies for cognitive radio users and we will propose an additional learning process on the primary users' side.

Acknowledgments

This work was partly supported by the Ministry of Higher Education and Research of France.

References

[1] J. Mitola, "Cognitive radio architecture: The Engineering Foundations of Radio XMLLink", John Wiley and Sons, 2006.

[2] Y. Tawk, J. Costantine, C.G Christodoulou, "Cognitive-radio and antenna functionalities: A tutorial [Wireless Corner]", *IEEE Antenna and Propagation Magazine*, vol. 56, n°1, pp.231-243, 2014.

[3] *K.*R Chowdhury, M. Di Felice, and I.F. Akyildiz, "Tp-crahn: a transport protocol for cognitive radio ad-hoc networks," *IEEE INFOCOM*, 2009, pp. 2482–2490.

[4] P. Ren, Y. Wang, Q. Duand, J. Xu, "A survey on dynamic spectrum access protocols for distributed cognitive wireless networks", *EURASIP Journal on Wireless Communications and Networking*, 2012.

[5] V. Krishna, Auction Theory. *Academic Press*, 2002.

[6] R. Myerson, Game theory: Analysis of conflict. Harvard University Press, 1997.

[7] Y. Zhang, C. Lee, D. Niyato, and P. Wang, "Auction approaches for resource allocation in wireless systems: A survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1020–1041, 2013.

[8] H. Bogucka, M. Parzy, P. Marques, J. Mwangoka, and T. Forde, "Secondary spectrum trading in tv white spaces," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 121–129, 2012.

[9] S. Maharjan, Y. Zhang, and S. Gjessing, "Economic approaches for cognitive radio networks: A survey," *Wireless Personal Communications*, vol. 57, no. 1, pp. 33–51, 2011.

[10] E. Tragos, S. Zeadally, A. Fragkiadakis, and V. Siris, "Spectrum assignment in cognitive radio networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1108–1135, 2013.

[11] X. Wang, Z. Li, P. Xu, Y.Xu, X. Gao, and H. Chen, "Spectrum Sharing in Cognitive Radio Networks- An Auction based Approach", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics - Special issue on game theory*, June 2010, vol. 40, n°3, pp. 587-596.

[12] L. Chen, S. Iellamo, M. Coupechoux, P. Godlewski, "An Auction Framework for Spectrum Allocation with Interference Constraint in Cognitive Radio Networks", *IEEE INFOCOM*, 2010, pp.1-9.

[13] Z. Chen and R-C Qiu, "Q-Learning Based Bidding Algorithm for Spectrum Auction in Cognitive Radio", *IEEE Southeastcon*, 2011, pp.409- 412.

[14] H. Song1, X-L. Lin, "An auction-based MAC protocol for cognitive radio networks", *International Journal of Communication systems*, 2012 vol. 25, n° 12, pp. 1530–1549

[15] H.-B. Chang and K.-C. Chen, "Auction-based spectrum management of cognitive radio networks," *IEEE Transactions on Vehicular Technology*, 2010, vol. 59, n°4, pp. 1923-1935.

[16] Y. Teng, F. Richard Yu, K. Han, Y. Wei, Y. Zhang, "Reinforcement-Learning-based Double auction Design for Dynamic Spectrum Access in Cognitive Radio Networks", *Wireless Personal Communications*, vol 69, n°2, 2013, pp.771-791.

[17] Y. Zhang, C. Lee, D. Niyato, and P. Wang, "Auction Approaches for Resource Allocation in Wireless Systems: A Survey", *IEEE Communications surveys & tutorials", vol. 15,n°3,* 2012. pp. 1020-1041.

[18] G. Pongor, "OMNeT: Objective Modular network Testbed", *International Workshop on Medeling, Analysis and Simulation On Computer and Telecommunication Systems* 1993, pp. 323-326.

Authors

Emna Trigui Doctor of networks, knowledge and organization from the University of Technology of Troyes (UTT, France); she received her engineering degree in Computer Engineering and her M.S. degree in Protocols, Networks and Multimedia Systems from the National School of Computer Science (ENSI, Tunisia). Her research interests include computer networks, wireless communication, mobile ad hoc networking and cognitive radio networks.

Moez Esseghir received the National Engineer Diploma in computer sciences from Ecole Nationale des Sciences Informatique (ENSI), Tunisia in 2002, the M.S. degree in networks from the University of Paris VI, France in 2003, the M.S. degree in computer sciences from ENSI, in 2004 and the Ph.D. degree in computer sciences from University of Paris VI in 2007. In 2008, he was with the North Carolina State University, Raleigh, USA, as Postdoctoral Fellow. Since September 2008, he is an associate professor in the Technology University of Troyes (UTT), France. His research interests include vehicular communications, wireless sensor networks, and spectrum sharing in cognitive radio networks.

Leila Merghem Boulahia received an engineering degree in computer science from the University of Sétif, Algeria, in 1998, an M.S. degree in artificial intelligence and a Ph.D. in computer science from the University of Paris 6, France, in 2000 and 2003, respectively. She is an associate professor at the University of Technology of Troyes (UTT) in France still 2005. She received the best paper award of the IFIP WMNC'2009. Her main research topics include multi-agent systems, quality of service management, autonomic networks, cognitive and sensor networks and smart-grids. Dr. Merghem-Boulahia authored or co-authored five book chapters, thirteen peer-reviewed international journal articles and more than 40 peer-reviewed conference papers. She also acted or still acts as TPC member of the following conferences and workshops (IEEE ICC, IEEE Globecom, IEEE GIIS, ACM/IEEE ICCVE, IFIP Autonomic Networking, IFIP NetCon…). She is a Member of IEEE.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Writer Identity Recognition and Confirmation Using Persian Handwritten Texts

Aida Sheikh [1*], Hassan Khotanlou [2*]

1. Department of Computer, Qazvin Branch , Islamic Azad university, Qazvin, Iran.

Aida.sheikh@gmail.com

2. Department of Computer, Bu-Ali Sina University, Hamedan, Iran.

Hassan.khotanlou@gmail.com

Abstract

There are many ways to recognize the identity of individuals and authenticate them. Recognition and authentication of individuals with the help of their handwriting is regarded as a research topic in recent years. It is widely used in the field of security, legal, access control and financial activities. This article tries to examines the identification and authentication of individuals in Persian (Farsi) handwritten texts so that the identity of the author can be determined with a handwritten text. The proposed system for recognizing the identity of the author in this study can be divided into two main parts: one part is intended for training and the other for testing. To assess the performance of introduced characteristics, the Hidden Markov Model is used as the classifier; thus, a model is defined for each angular characteristic. The defined angular models are connected by a specific chain network to form a comprehensive database for classification. This database is then used to determine and authenticate the author.

Keywords:*Persian handwriting recognition, authentication, Off-line, Hidden Markov Model*

1. Introduction

In this era, security and information protection is considered a big challenge of humanity in the modern world. Identifying writers from their handwriting has recently become a considerable and interesting subject in identity recognition. Among behavioral features, handwriting is easily achieved and studies show that different individuals have different handwritings [1,2]. Therefore, identifying and confirming the identity of individuals using their handwriting has been a research focus in recent years and its major application is in security and legal issues, controlling systems access, and financial activities. The identity recognition problem aims to specify the identity of the writer, given a handwritten text and the identity confirmation problem aims to specify whether two given handwritten texts have been written by one individual. Generally, graphology experts analyze and investigate handwritten texts. Although human intervention in solving this problem is an effective approach, it is costly and tiring due to the nature of human beings.

Most conducted studies in writer identity recognition and confirmation have been offline and signature-based online methods are more common in identity recognition. Most studies in identity recognition focus on the English language and there has been few studies regarding Arabic and Persian handwritten texts in comparison to the English language. Accordingly, this paper briefly studies the widely used methods in Latin and Persian texts and compares these methods and application for the Persian language. Finally, an online method is introduced and presented for identity recognition and confirmation for Persian texts.

2. Previous Works

Reviewing conducted studies and papers shows that they mostly investigate writer identity recognition based on handwritten texts and a limited number of methods are introduced for determining and confirming identity. This may be because introducing an efficient solution for identity recognition can also include a reliable approach for identity confirmation. In what follows, the most important proposed methods are introduced in the domain of handwriting recognition.

Horizontal and vertical views have been widely used in identifying English and Persian texts. Accordingly, Zois et al. [3] propose a text dependent method for identity recognition. In this method, the vertical view histogram of a word`s image is processed and the considered features are extracted using morphology operators.

Bensifa et al. [4] investigate identity recognition in form of a text-based information retrieval problem. In this method, features are retrieved based on the type of the used data and thus, this method is easily generalized to other languages, including Arabic, Persian, etc.

Reference [5] proposes a method for identity recognition and confirmation, which is based on Markov`s hidden model. In this model, for each writer, an HMM model is considered and each corresponding model is trained using samples of the writer`s handwriting. Therefore, each HMM model works as an

expert, who is able to recognize the handwriting of a certain individual.

Reference [6] introduces a method for writer identity recognition that extracts features from the image of the handwritten text using edges information. The introduced features indicate direction changes during the writing process. In fact, the main notion of this method is that the writing process constitutes of a set of pen hits and the specifications of this process can be achieved by extracting the edges information.

In a paper by author [7], a new method is proposed for feature extraction from handwritings. One of the proposed methods is the n feature extraction method that extracts features indicating straight lines and curves used in handwritings.

The aforementioned methods are the basic propose algorithms in handwriting recognition. Table 1 briefly presents some of the writer identity recognition papers using different algorithms.

Tabe1: writer identity recognition papers using different algorithms

Author's Name	***Database***	***Feature Extraction***	***Classification Methods***	***Language-Accuracy***
M.Bulacu [8]	650 Writers	Edge based directional PDFs as features (Textural and allograph prototype approach)	K-nearest neighbor and a feed forward neural network classifier	English- 92%
Neils et al. [9]	43 Writers	Allograph prototype matching approach usingthe dynamic time warping(DTW)distance function	af-iwf (allograph frequency inverse writer frequency) measure	English- 60%
B.Helli,al. [10]	100 Writers	Point-based (speed,acceleration, vicinitylinearity, vicinity slope), stroke-based	Tey proposed an LCS (longest common subsequence) based classifier	Persian- 95%
Soleymani Baghshah [11]	128 Writers	Fuzzy approach	Fuzzy learning vector quantization	Persian- 95%
Helli,B.,Moghaddam [12]	100 Writers	XGabor filter	Weighted Euclidian distance	Persian-77%
Sadeghi [13]	50 Writers	Fuzzy clustering & gradient features	MLP(multi layer perceptron)	Persian-77%
Golnaz Ghiasi [14]	180 Writers	Codebook	2-fold cross validation, Leave-oneout cross validation	Persian-93%
Schlapbach et al. [15]	100 Writers	X-Y coordinates	Hidden Markov Models	English- 96%
Bensefia et al. [16]	88 French Writers 150 English Writers	A textual based Information Retrieval model, local features such as graphemes extracted rom thesegmentation of cursive handwriting	Cosine similarity	French- 95% English- 86%
Rafiee and Motavalli [17]	100 Writers	Baseline and width structural features	Feed forward neural network	Persian-95%

3. The Structure of the Recognition System

Handwriting as a behavioral feature is very dependent on brain performance and it is affected by several complex factors, like training, physical status, environmental and mental conditions of the individual. Therefore, in writer identity recognition, features should be discussed that indicate the differences between

different handwritings. In addition, Arabic and Persian handwritings have several differences from Latin handwriting. Persian and Arabic handwritings are inherently continuous and overheads and underlines prevent the application of methods that require full segmentation. The shapes of Persian and Arabic alphabets are a function of their location in the word and there may be different forms of each alphabet in different locations, including the beginning, middle, and ending of the word or as a separate alphabet. Some Persian alphabets have one, two, or three dots above or below them. Persian and Arabic handwritings have many print fonts and handwriting styles.

3.1. The Proposed Algorithm

Since our goal is to propose an automatic writer identity recognition method and there are no constraints in the considered handwritings, we cannot use methods that require automatic and full segmentation of text into words and alphabets. The general structure of this method includes four stages: collecting the handwriting samples, preprocessing and create the binary image, feature extraction, writer identity recognition or confirmation using pattern recognition methods. The proposed method considered the handwritten text as an image. This method is evaluated using the handwritings of 70 individuals and results indicate that the proposed method is highly efficient for Persian handwritten texts.

3.1.1 Sample Collection

In order to evaluate the proposed method and the other investigated approaches, 70 individuals with different education, age, and gender were asked to fill the designed forms with their normal handwriting. These individuals were selecting from ordinary people. For evaluation, the individuals were asked to write two determined texts, which are in one paragraph and include various words. The determined texts are presented in figure 1 and 2.

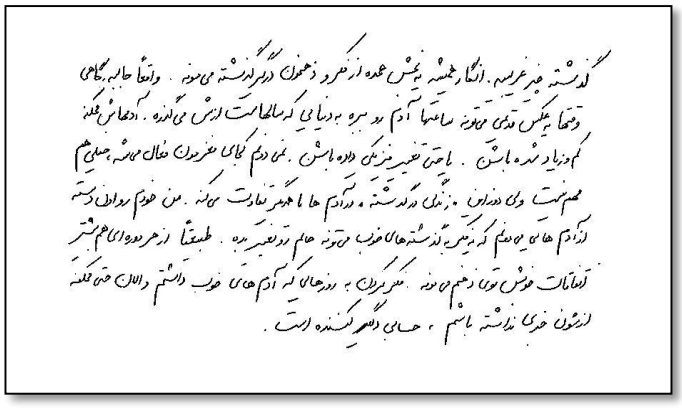


Figure 1. The image of the first handwritten text

The forms do not have lines; however, the individuals were asked to write along straight horizontal lines. Nonetheless, it was observed was some forms had tilted lines; in most forms, however, the lines were almost align a straight line. After collecting the forms, all were scanned as grayscale images with resolution 300dpi.

3.1.2 Preprocessing and create binary images

One of the stages that is performed in most image processing systems is binarization. Transforming a grayscale image into a binary one has several advantages. The size of a binary image is far smaller than a grayscale one and since, one of the important characteristics of personality recognition systems for handwritten texts is their speed, and the smaller size of the binary image facilitates and accelerates all stages. Therefore, create binary image reduces image sizes, as well as removing redundant information.

3.1.3 Feature Extraction

In order to propose an efficient method for writer identity recognition, features should be considered that indicate the differences of different handwritings. in this method, angular features (in 8 angles, 0, 20, 40, 60, 80, 100, 120, and 140 degrees) are first extracted and compared for each image page.

I(x, y) is the resulting image from preprocessing, which is a binary image. First, closing morphological operations are applied to binary image I. $g_{i:}$ is the i-th image resulting from closing operations with constructing element S_i, which has the same dimensions as the input image I. I: the binarized input image. S_i: the i-th component. Components are direct lines with different angles. Figures 3 to 5 presents an example of angular specifications extracted from each image.

As we can see, for most angles, these specifications have different patterns for different images and by inferring these differences, features can be introduced that can recognize the handwriting. Some of the usable features include 1- the number of remaining black objects in the image resulting from angle extraction operation. For instance for a 40 degrees angle, it is clear that the image has more black spots. 2- The length of the remaining objects from the feature extraction operation. 3- The total area of black locations of the image resulting from the feature extraction operation.

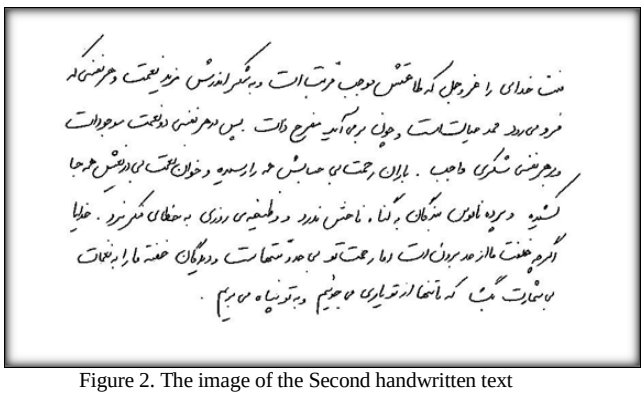


Figure 2. The image of the Second handwritten text

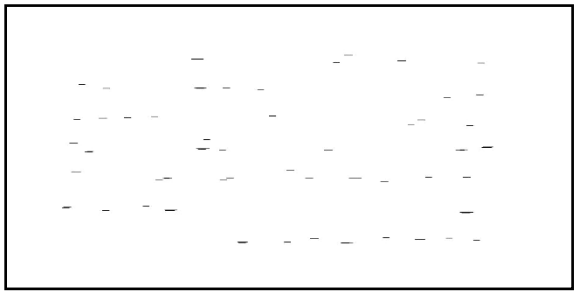


Figure 3. Angular specifications of 0 degrees for the image of a handwriting

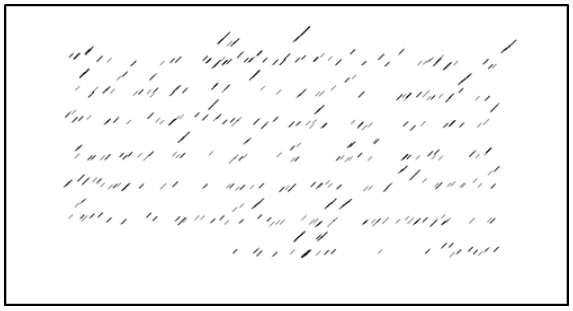


Figure 4. Angular specifications of 40 degrees for the image of a handwriting

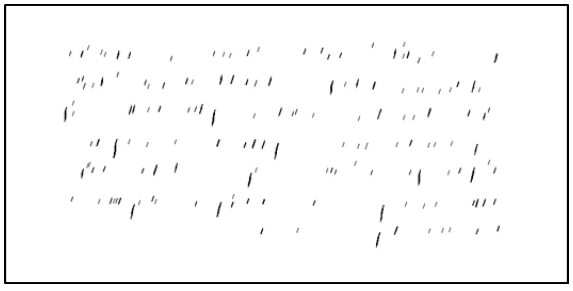


Figure 5. Angular specifications of 80 degrees for the image of a handwriting

Subsequently, the angular specification of each image is extracted for different angles of the number 2 handwriting of the collected samples as figure 2. Moreover, we can also use a combination of the aforementioned features. For the feature extraction stage, the number of length of the remaining black objects is used as the component. This feature is computed and compared for the handwritings of two individuals that there are two different handwritings from each individual. As we can see in table 2, the difference of features between the handwritings of an individual is smaller than that of two different individuals.

After the pre-processing stage, feature extraction is performed for each image extracted from angular specifications, separately with components in form of lines with different lengths, which are considered 3 to 40 in this research, so that the observation sequence (feature vector) of each handwritten text is achieved. Finally, this observation sequence is sent to a chain network to recognize the identity of the writer; however, it should be normalized before sending it to the classification stage. Moreover, when the discrete hidden Markov model is used for classification, the input values should be discrete. Therefore, a quantization operation is used. A feature vector with integer values in range [0, 100] is considered in the implemented identity recognition system for each angular specification.

Table 2. Calculating features for the handwritings of two individuals

Angles	*Number of components remained from First Handwriting of first person*	*Number of components remained from Second Handwriting of first person*	*Number of components remained from First Handwriting of Second person*	*Number of components remained from First Handwriting of Second person*
0	34	50	64	74
20	166	163	147	159
40	199	214	146	166
60	104	101	35	40
80	83	79	29	33
100	25	23	13	11
120	5	7	4	7
140	15	15	10	8

4. Proposing a Model for Angular Specification

The proposed writer identity recognition system is divided into two main sections used for training and testing. In order to increase accuracy and security, the constructed models are embedded in a chain network, which presents a credible text image for angular specifications. In this research, each angular specification is considered as a separate model. The discrete left to right hidden Markov model [18] with a forwarding state is used to implement each angular specification model. In order to train the model, it is necessary to compute the related parameters, $\lambda = \{A, B, \pi\}$, using equations 1, 2, and 3. Therefore, it is only necessary for the considered angular specification after the feature extraction stage and providing the feature vector (observation sequence o_i) to use the three equations above to extract the parameters of the model.

$$\bar{\pi}_i - \gamma_1(i), 1 \le i \le N \quad (1)$$

$$\bar{a}_{i,j} = \frac{\sum_{t=1}^{T=1} \xi_t(i,j)}{\sum_{t=1}^{T=1} \gamma_t(i,j)}, 1 \le i \le N \quad 1 \le j \le N \quad (2)$$

$$\bar{b}_j(k) = \frac{\sum^{T}_{t=1,\ o_t=v_k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, 1 \le j, \quad 1 \le k \le M \quad (3)$$

4.1 Feature Extraction

The Baum Welch algorithm is used to train HMM parameters [18]. It means that $\lambda = \{A, B, \pi\}$ is used to present each angular specification model. The initial state distribution of the models are $\pi = \{\pi_i\}$, where $\pi_1 = 1$ and $\pi_i = 0$ for $1 < i \le N$. Moreover, N is the number of states in the corresponding model. For the transition

probability matrix $A = \{a_{i,j}\}$, initially $a_{i,j} = a_{i,i+1} = a_{i,i+2} = 0.5$ for $1 \leq i < N$ and $a_{i,j} = 0$ where $j \neq i + 1$, $j \neq i$, and $j \neq i + 2$. Moreover, $a_{N,N}$ is equal to one.

4.2 Proposing a Chain Network for Training Models

A chain network of angular specifications is used to increase the accuracy and security of the writer identity recognition system`s output and constrain the search space. The chain network is divided into a set of sub-models $\{D_1, D_2, \dots, D_n\}$ based on the angles. Subsequently, each model is indexed as $SWD_{k,i}$ based on its parts. Sub-models are embedded in a chain network called $WN_{k,i}$, which presents the text images, and they are connected through specific paths. The performance of the identity recognition system for searching a writer on the chain network is as follows. First, using the method, which was explained in the feature extraction method, the observation sequence $O_s = [o_1, o_2, \dots, o_k]$ is obtained for the image of a handwritten text.

In this equation, $p(SW_i|O_i) = p(O_i|SW_i)p(SW_i)/p(O_i)$ [18]. Moreover, for simplicity, we assume that all parts of the text image occur with the same probability along the chains` path. Therefore, $p(O_i)$ is the same for all parts of the image text and the problem is reduced to maximizing$p(O_i|SW_i)$, which can be effectively compuited by the Viterbi Algorithm for network $WN_{k,i}$. Finally, equation 4 is used to search writer w in D_k.

$$W = \text{argmax} \quad \prod_{i=1}^{k} P(O_i|SW_i, WN_{k,i}) \qquad (4)$$

5. Implementation of the Proposed Algorithm

In this research, all simulation is performed using Matlab R2013a on windows 7 operating system and 2.30GHz Intel Core i5 processor. Results of extracting the components based on angular specifications show that for Persian handwritten texts, features are about the same throughout the text. For instance, for a 40 degrees angle, images with maximum component have different pixel lengths. After the pre-processing stage, for each image extracted separately from the angular specification with line components and different pixel lengths, which is considered 3 to 40 in this research, feature extraction is performed to provide the observation sequence (feature vector) corresponding to each handwritten text.

Subsequently, for each writer, a left to right HMM is considered in which the initial node is zero degrees and the end node is 140 degrees with a 20 degrees step. After the feature extraction stage, for the first handwriting of each writer, the state change probability matric and symbol observation probability matrix is created, which is considered the input of the HMM. Parameters are trained using the Baum–Welch algorithm on the extracted features. In these computations, the maximum learning iteration is considered 100. In order to classify a sequence to one of the K classes, K HHM models are trained for each class. Subsequently, LL values are computed using the HMM Toolbox [18] of MIT. If the i-th model is the most similar, the sequence belongs to class i. in statistics, the log likelihood function is a function of statistical model`s parameters that plays a key role in inferential statistics.

In order to run the propose algorithm to recognize and confirm the identity of the writer, a specific handwriting of a writer is selected; the second handwriting of the same writer is then compared using HMM Toolbox. Subsequently, the second handwriting of another writer is selected and its LL measure is computed. In this section, the output of the proposed identity recognition system is compared under different circumstances. In these experiments, from the total 70 samples, 50 are selected for training and the remaining 20 are selected for testing. This comparison is performed for different training conditions, including the parameters of components` angles, training duration, etc. the precision is computed based on the percentage of correctly recognized samples divided by the total number of recognitions. Table 3 presents these results. As we can see, the highest precision rate of identity recognition for the training data is resulted when the number of component`s angle changes is equal to 7 or 12 and the component length is equal to 9 or 10. Whereas, for the test data, the number of component`s angle changes is equal to 12 and the component length is equal to 9.

Table 3. Comparison of results for different HHM training conditions

number of component`s angle changes ($\Delta\theta$)	*component length*	*Learning Time (Second)*	*precision rate of identity recognition*	
			Learning Data	*Test Data*
4	17	804	99%	63%
7	13	444	98%	75%
10	10	320	98%	78%
12	9	266	94%	81%
15	8	213	89%	77%
17	7	177	84%	74%
20	6	160	78%	73%
25	5	124	71%	65%
30	4	106	61%	57%

6. Evaluation

Since, there has been no standard database for Persian handwriting texts and the relevant published papers each have a different database, the Arabic handwriting database in reference [20] is used to evaluate the proposed algorithm.

This database consists of 4 pages in Arabic of 1000 forms of handwritings, which are collected from 1000 female and male writers in the country with different ages, 15 to 70. About 65% of these forms are filled by individuals with ages in range 16 to 25. For each individual, there are four images of handwritings with

different contexts in resolutions 300 and 600dpi. Reading signs and symbols are rarely observed in these texts.

After scanning the forms and performing pre-processing operations, the researchers, who have collected this database [19], have transformed them into binary images using the Otsu thresholding algorithm and they have also removed the noise. The tilt modification operations are performed before extracting paragraphs from the images. After identifying the possible lines, the X-Y coordination of each extracted line and their regression is obtained to compute their slopes. At the next stage, paragraphs are extracted. Since, there are many dots and overheads between the lines in Arabic, the subsidiary components are first removed and only the primary components remain (in the reference paper, this part is called Part of Arabic Word (PAW)). The image is then divided into several vertical strips. The lines of each strip are identified and finally, dots and overheads are added to the strips. The words of each line are identified by finding the white spaces between words. Moreover, the division positions are identified by findings a hypothetical space more than 10 pixels. At the next stage, the text inserted in the forms is evaluated in three phases. This is performed by human personnel. In the performed research, results of 4808, 938, and 966 lines are respectively dedicated to training, validation, and test sets. The handwritten text is recognized using discrete left to right HMM with Bakis topology and using HTK. In fact, the shape of each character is considered as a class during training. The identified characters with similar shapes are placed in one class. In sum, 149 classes are considered. Several statistics features are extracted from the lines images using the sliding window technique, where windows are considered with different lengths. However, the best result is achieved by the sliding window with length 4 and shared space 2. After the extraction, features are quantized in the linear codebook using top-down clustering. Table 4 presents the recognition rate results of the algorithm in reference [19], the multi-layer perceptron neural network, and the proposed algorithm.

Table 4. Comparison of the precision results of the proposed algorithm on the handwriting database

Methods	***Training***	***Test***
	Acc. (%)	***Acc. (%)***
Reference [19]	51.4	51.73
MLP	28	34
Proposed algorithm	61.5	58.7

Since this research aimed to propose a dynamic (smooth and automatic) method and no limitations are considered for the type of the investigated handwritings, we do not consider methods that require the automatic segmentation of the text into words and letters. The characteristic of the proposed method is compatible with the structure of Persian handwritten texts.

This method is introduced as an approach to text-dependent identity recognition. Of course, considering the images resulting from angular specifications, it is clear that these features are about the same throughout the text and thus, this method can also be used as an independent method.

7. Conclusions

Feature extraction is one of the determining stages of increasing the identity recognition rate in identity recognition systems. Extracted features should indicate the different of an individual`s handwriting from that of others and there should also be minimal difference between handwritten texts of an individual. This research employs a combination of edge direction distribution and edge axis distribution features to extract features, as well as the hidden Markov Model. Angular specification models are connected through a specific chain network to increase the accuracy and security of the identity recognition system. Moreover, a database provided by 70 different writers was used to train and evaluate the system.

Considering the results, the advantage of the proposed algorithm is as follows:

1- Using a combination of edge direction distribution and edge axis distribution features at the feature extraction stage.

2- The autonomy of the proposed identity recognition system, particularly at the feature extraction stage (since, manual feature extraction is overwhelming and time-consuming).

3- Increasing the accuracy and credibility of the identity recognition system due to using the chain network at the training stage.

4- Eliminating the need for the segmentation stage at the feature extraction stage.

Since the writer identity recognition problem is performed for the first time using Persian handwritten texts with the proposed feature extraction method, the results of this paper are promising. Moreover, the proposed features are global features and the results can be improved by extracting and combining different features.

References

[1] S.N. Srihari, H.Arora, S.H. Cha and S.lee,"Individuality of handwriting," Journal of Forensic Sciences, vol.47, no.4, pp.1-17, 2002.

[2] S.N. Srihari, H.Arora, S.H. Cha and S.lee,"Individuality of handwriting: a validation study," IEEE Proc. Of 6th Int. conf. on Document Analysis and Recognition, pp.106-109, 2001.

[3] E.N.Zois,V.Anastassopoulos, "Morphological waveform coding for writer identification," Pattern Recognition, vol.33, pp.385-398, 2000.

[4] A.Bensifa, T.Paquet and L. Heutte, "Handwritten document analysis for automatic writer recognition," Electronic Letters on Computer Vision and Image Analysis, vol.5, no.2, pp.72-86, 2005.

[5] A.Schlapbach and H.Bunke, "Using HMM-based recognition for writer identification and verification," IEEE Proc of ٩th Int. Workshop on frontiersin handwriting Recognition, pp.167-172, 2004.

[6] M. Bulacu, L.Schomaker, "Writer style from oriented edge fragments," proc.10th Int. Conf. Computer Analysis of images and Patterns, pp. 460-469, 2003.

[7] Golnaz Ghiasi, Reza Safabakhsh, "An Efficient Method for Offline Text Independent Writer Identification" International Conference on Pattern Recognition, 2010.

[8] Bulacu, M., Schomaker, L., Vuurpijl, L. Writer identification using edge-based directional features, in: Seventh International Conference on Document Analysis and Recognition (ICDAR).2005.

[9] Niels, R., Gootjen, F. Vuurpijl, L. "Writer Identification through Information Retrieval: TheAllograph Weight Vector," in International Conference on Frontiers inHandwriting Recognition, pp. 481-486, 2008.

[10] B. Helli, M.E.Moghaddam, "A text-independentPersian writer identification system using LCS based classifier", in: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2008.

[11] Soleymani Baghshah, M., Bagheri Shouraki, S. Kasaei, S. a novel fuzzy classifier using fuzzy LVQ to recognize online Persian handwriting, in: Second IEEE Conference on Information and Communication Technology (ICTTA), 2006.

[12] Helli, B.Moghaddam, M.E. Persian writer identification using extended Gabor filter, in: International Conference on Image Analysis and Recognition (ICIAR), 2008.

[13] R.Sadeghi .S.Moghaddam, M.E. Text-independent Persian writer identification using fuzzy clustering Approach, in: International Con- ference on Information Management and Engineering (ICIME), Malaysia.2009.

[14] Golnaz Ghiasi, Reza Safabakhsh, "An Efficient Method for Offline Text Independent Writer Identification" International Conference on Pattern Recognition, 2010.

[15] Schlapbach, A., Bunke, H. Using HMM based recognizers for writer identification and verification, in: Proceedings–International Workshop on Frontiers in Handwriting Recognition, IWFHR, Tokyo, pp. 167–172, 2004.

[16] A. Bensifa, T. Paquet and L. Heutte, "A Writer identification and verification system," Pattern Recognition Latters, vol.26, no.13, pp.2080-2092, 2005.

[17] Rafiee, A., Motavalli, H. Off-Line Writer Recognition for Farsi text. In: 6th Mexican International Conference on Artificial Intelligence, Special Session,pp. 193-197, 2007.

[18]http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm_usage.html

[19] KHATT: S. A. Mahmoud, I. Ahmad, W.Al-Khatib, M. Alshayeb,M. Tanvir Parvez, V. Märgner, G.A. Fink, "An open Arabic offline handwritten text database", Pattern Recognition, vol. 47, no. 3, pp. 1096–1112, March 2014.

[20] http://khatt.ideas2serve.net/index.php

Aida Sheikh received her B.Sc. in computer engineering from department of computer Engineering, Hamedan Branch, Islamic Azad University, in 2007, and her M.Sc. in computer software engineering from department of computer, Qazvin Branch, Islamic Azad University, in 2015.Her research interests includes Image processing, knowledge management and intelligent systems.

Hassan Khotanlou is an associate professor in department of computer at Bu-Ali Sina University, Hamedan, Iran. He received his B.Sc. degree in computer engineering from IUST University in 1995, and his MSc. degrees in artificial intelligence engineering from Shiraz University in 1997 and his Ph.D. degrees in artificial intelligence engineering from Pierre & Marie Curie University (Paris VI – TELECO) in 2007.His main research interests are Image processing, Fuzzy Systems, Acoustic, Data Mining, Computer Networks, Medical Image processing and Artificial Intelligence.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Detecting features of human personality based on handwriting using learning algorithms

Behnam Fallah[1], Hassan Khotanlou[2]

[1]Department of Computer, Qazvin Branch, Islamic Azad University, Qazvin, Iran.

bmfallah@yahoo.com

[2]Department of Computer, Bu-Ali Sina University, Hamedan, Iran.

Hassan.khotanlou@gmail.com

Abstract

Handwriting analysis is useful for understanding the personality characteristics through the patterns created by the handwriting and can reveal features such as mental and emotional instability. On the other hand, it is difficult to determine the personality, especially when it is associated with the law because there is no threshold or scale being able to make detailed results of the analysis. This thesis aims to provide an automated solution to recognize the personality of the author by combining image processing and pattern recognition techniques. The personality recognition system proposed in this project is composed of two main parts: training and testing. In the training part, after feature extraction from all image patterns of the input text, a proportional output is created through the MMPI personality test. Then these inputs are trained to the neural network as a pattern. As a result of this training, a comprehensive database will be formed. In the testing part, the database is used as a main comparison reference. After feature extraction, the input text image is compared with all patterns in the database to find the closest image to the input text image. Finally, the MMPI personality test output for the proposed text image is introduced as the output personality parameters.

Keywords: *handwriting recognition, neural network, MMPI personality test, graphology*

1. Introduction

Handwriting analysis is an act that has been performed for many years. However, when we analyze the behavior and personality of an individual, its effects are still discussable. Each of these neural brain patterns lead to a unique neural and muscular movement for an individual; therefore, this small subconscious movement occurs for each person, who has a certain personality feature while writing. the problem of recognizing the writer`s personality from his handwritten texts aims to specify the personality of the writer given a handwritten text, which of course this personality elicitation is performed based on samples of different individuals` handwritings. Professional handwriting researchers investigate and analyze the sample handwriting.

Writing can indicate personality features like feelings, fear, honesty, etc. identifying the personality of a human being by his handwriting is an old technique. Before, the nature of an individual was predicted manually, which took a long time. Recognizing a writer`s personality from his handwriting has recently become a considerable and interesting subject in psychology.

2. Previous Works

This section discusses a number of feature extraction methods from Persian and Arabic alphabets, words, and handwritten texts.

Reference [1] reviews identity recognition methods from handwriting and graphology. Moreover, a review of handwriting analysis computer systems in the market is presented and compared for better understanding.

Reference [2] points out and uses salient and important features of using handwriting in graphology analysis, including curves and figures in paper margins, line spaces, line tilts, word tilts, sharpness of edges, character sizes, text density, writing speed, and order point. This paper proposed some methods for the first time. Moreover, for automatic feature extraction from Persian handwritten texts, 24 training samples and 118 test samples are used for the experiments.

Reference [3] considers 6 different feature types for computerizing handwritten graphology: 1- character sizes, 2- word tilts, 3- baseline, 4- pen pressure, 5- space between characters, and 6-space between words in the document that is used to recognize the writer`s personality.

Reference [4] claims that potential behavioral and personal deviations of an individual are possible by analyzing his handwriting. In this paper, two methods are proposed for handwriting analysis:

1- Graphology that is a psychological analysis method.
2- Graphology that is used to identify the writer.

Reference [5] discusses the effect of brain neural patterns on micro-movements of the muscles, such that because of these micro-movements, personal parameters are emerged in human daily behaviors like writing a text. All hits, patterns, and the pressure that is applied when writing a word can express personal behaviors of an individual. In this reference, features like the tilt of the baseline, pen pressure, tilt and size of characters are used to extract the personal parameters of an individual. This reference used the linear regression method at the classification stage.

Reference [6] extracts personal parameters like extrinsic emotions, fear, trustworthiness, etc. from the graphology of handwriting. Professional handwriting tests, which are called graphology, are mostly recognized by part of the handwritten text. The handwriting analysis accuracy depends on the skills of the expert. In this paper, a method was proposed to investigate the personality based on features like baseline, pen pressure, and character t in a separate handwritten text. These parameters and features are defined as the input of a neural network, whose output is the personal behavior of the writer. The performance of this system is measured by different experiments.

Reference [7] uses computers to accelerate the image processing of Persian handwritten texts. More specifically, important features are extracted from the handwritten text to understand the psychological state of the writer. In order to do so, features like boldness, compression, and two word space measures are achieved. The tilt angle of the text is then obtained and after removing it, the main points corresponding to the right, left, and up margins, as well as the shape of the margins are achieved. Finally, the line vibration is obtained. Extracted features are divided into four groups, which are sent to three fuzzy systems and a non-fuzzy system to recognize the personality of the writer.

3. General Principles of Personality Recognition System from Persian Handwritten Texts

Figure 1 presents the block diagram of a personality recognition system from the handwritten texts. The innovative methods in personality recognition from handwritten texts are ensued by changes in running one or several blocks of the following diagram. The general name of the system is in fact changed to the block performance. For instance, recognizing personality from handwritten texts based on neural networks is in the classification section.

3.1. Pre-Processing

All operations performed on texts, which facilitate the process of following phases, e.g. removing noise, smoothing, thinning, language recognition, word fonts, etc. the set of these processes have the following goals [8]:

1- Reducing noise

2- Contour smoothing

3- Storing information that should be protected (thinning)

Activate binarizing for facilitating working with edit, ruler, and paragraph markers. By activating “Ruler” in the “View" section, you can see the settings of specific distances, columns, and margins. In order to find the undesired parts of paragraphs, like spaces between alphabets, pages, and lines, specific spaces, and chapter headers, activate paragraph markers (¶) in the “Paragraph” toolset.

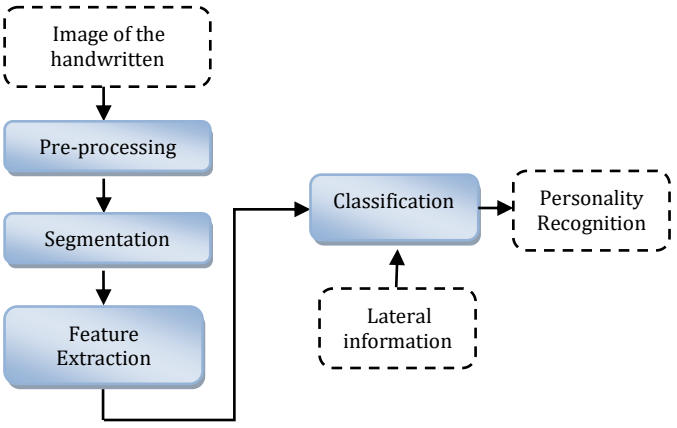


Figure 1. The block diagram of a personality recognition system

3.2 Segmentation (Separation)

The segmentation stage is very important for feature extraction, and calculating the sizes of characters and words in personality recognition systems from handwritten texts [9].

3.3 Feature Extraction

Feature extraction aim is to reduce data to a set of features, such that they are constant in the handwritten texts of a certain individual and independent of the other person's handwriting.

3.3.1 Text Independent Features

Independent text features include the margin value from the beginning of the page, word expansion, characters sizes, line spaces, word spaces, word tilts, horizontal to vertical ratio of characters, and lie tilts.

3.3.2 High-Order Local Autocorrelation as a Text Dependent Feature

Autocorrelation is one of the most well0-known functions insensitive to shift [10]. In what follows, a special type of autocorrelation is introduced:

Autocorrelation function is known as a shift insensitive function. The N-th order autocorrelation

function is defined by N location changes$(a_1, a_2, a_3, \dots, a_N)$, from reference point r as equation (1).

$$x^N(a_1, a_2, a_3, \dots, a_N) = \int I(r)I(r + a_1) \dots I(r + a_N)dr \quad (1)$$

Where, I(r) is the text image and $r = z = (\log(p), \boldsymbol{P})$. There are many autocorrelation functions, which are resulted from different combinations of location changes on the text image. Figure 3 presents different masks with location change patterns obtained by the autocorrelation function. In this research, the number of location changes is reduced by removing location changes that are equal due to the same shift. In other words, the value of N is limited to 3 (N= 0, 1, 2, 3). Moreover, the ratio of location changes and displacements are limited in a 3×3 position window [10]. In order to obtain the values of HLAC, the text image is first scanned by 70 local 3×3 masks. The sum of the multiplication of corresponding pixels` values by one (pixels of black mask) is then computed for each mask [11]. This method is known as HLAC feature (figure 2).

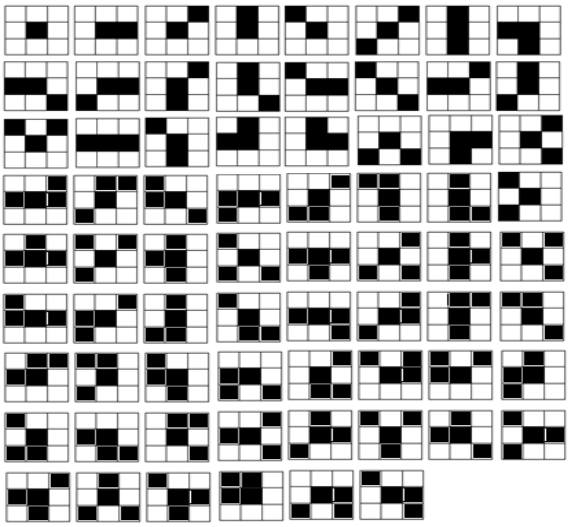


Figure (2). 70 masks with different patterns to extract HLAC [8].

3.4 Feature Vector and Generalized Discriminant Analysis (GDA)

In the proposed personality recognition system, when the number of training patters is greatly increased, the resolution of HLAC is reduced, which causes class interference. In this research, generalized discriminant analysis (GDA) is used to increase the space and resolution of different classes [12]. GDA is resulted from the non-linear extension of the linear discriminant analysis and it is successfully used for many applications. This method can be used to overcome classification problems. GDA helps us to combine features and increase the resolution of the classes.

3.5 Classification and Recognition (With One or Several Classifiers)

This stage includes methods to map each pattern extracted from the feature extraction stage with one of the classes of the corresponding pattern space. This is performed through minimizing the feature vector of each input pattern in proportion to one of the reference vectors. Reference vectors are vectors that are derived from the training samples a priori. The proposed techniques for this stage can be searched in one of the general pattern recognition discussion groups [12]:

- Pattern matching
- Statistical techniques
- Neural networks

The three groups above are not necessary performed separately and they may be found among the techniques of other groups.

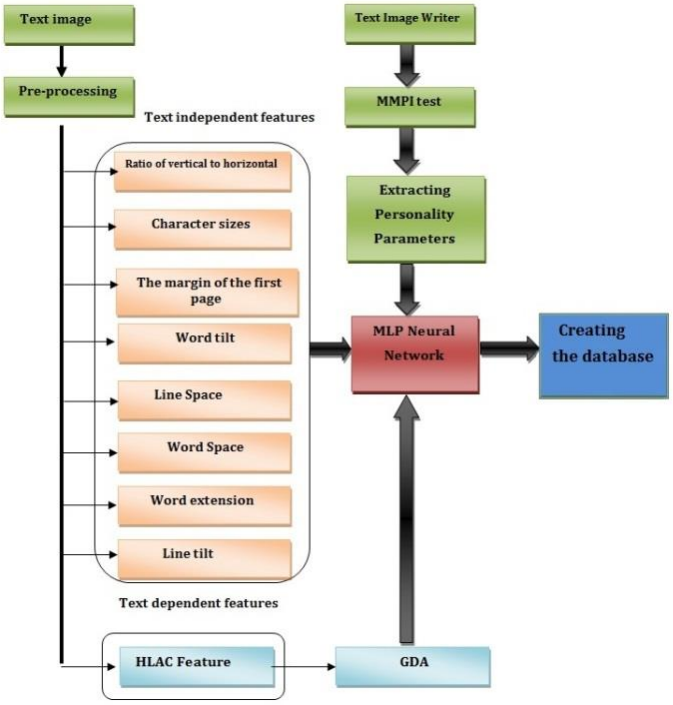


Figure 3(a). the block diagram of the personality recognition Training system

4. The Structure of the Handwriting Personality Recognition System

The proposed personality recognition system consists of two main training and test sections (figure 3). In the training section, after extracting features from all patterns of the input text image, the corresponding output is created using a MMPI personality test [13]. These input-outputs are then trained as a pattern in a neural network and finally, a comprehensive database is created as a result of training.

In the test section, this database is used as the main comparison reference. At this stage, after feature extraction, the input text image is compared with all patterns in the database to find the closest image. Finally, the output of the MMPI personality test corresponding to the selected text image is introduced as output personality parameters.

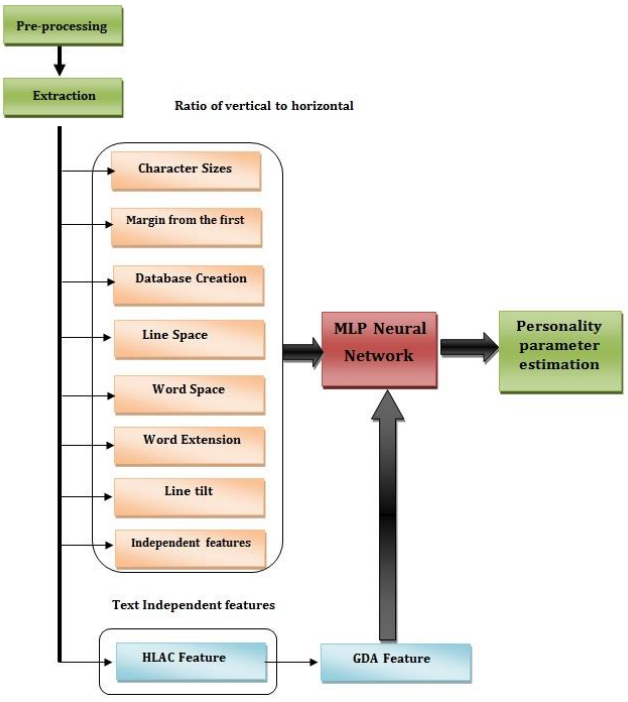


Figure (3b). the block diagram of the personality recognition Test system

5. Experimental Conditions and Database Type

In order to evaluate the proposed method, 70 individuals with different educations, ages, and genders were asked to fill the designed forms with their ordinary handwriting. These individuals were selected from ordinary people (college students, employees, etc.). Table (1) presents the characteristics of these individuals, including gender, education, and age.

Table 1. The characteristics of the individuals who filled the database's forms

Gender		Education			Age	
Female	Male	Elementary school to diploma	Diploma to undergraduate	Undergraduate to doctorate	7 to 40	Higher than 40
32	38	18	36	16	37	33

In order to perform the experiments and evaluations, these individuals were asked to write a constant text in one paragraph, which contained various words.

Subsequently, each writer took the MMPI personality test, which was designed in 1940s by two researchers of the University of Minnesota, United States. This personality test consists of 71 fundamental questions and the writer should select from true or false options. Finally, a score is given to the test that is in fact the character profile based on eleven clinical scales. These results are stored together with the ID that is assigned to each writer before writing the corresponding paragraphs.

The following points are considered when filling the forms:

- There is no limitation regarding the handwriting type. The collected samples include various handwritings. Of course, in some cases of writings filled by female individuals, the handwritings are very similar.
- The individuals were asked to fill the forms in a specific time duration and write using their ordinary handwriting without trying to alter or improve it.
- The forms do not have lines and only the paragraph area is specified.

Since the handwriting of each individual is changed by his/her mental and environmental conditions, the subjects were asked to fill the forms with patience and tranquility. After collecting the forms, they were all scanned as gray-scale images in 300dpi resolution.

6. The Environment of the Simulated Software

In this dissertation, all simulation are performed using MATLAB 2013a and in some cases, Matlab instructions are used to transform gray-scale images into binary ones and label groups. Moreover, DRTools[1] toolbox is used to apply the GDA algorithm, whose reception and route should be separately defined in the Matlab toolbox.

7. Results of the Proposed Algorithm

In this section, the output of the proposed personality recognition system is compared with those of MMPI test (i.e. the 11 scales of the MMPI test). Figure 4 presents the HLAC feature before and after using the GDA algorithm:

[1] Dimensionality Reduction

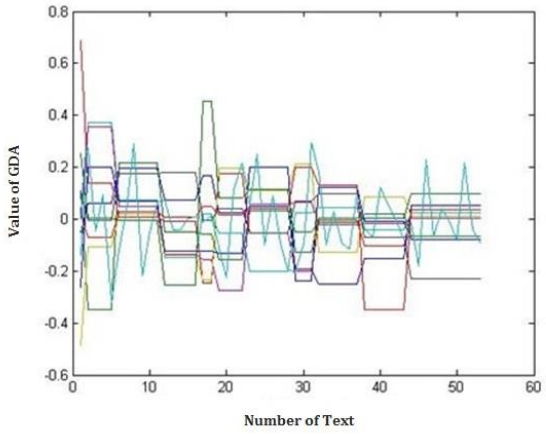


Figure 4. The extracted features for 53 handwritings without using GDA

In figure 4, the horizontal axis represents the number of handwritings and the vertical axis shows the value of the extracted feature. As we can see in the figure, after applying GDA, the resolution is considerably increased. In these experiments, from the total 70 samples, 50 were selected for training and the rest was used for tests. This comparison is performed for different training conditions, including the main parameter of the MLP neural network, e.g. the number of neurons in the middle and input layers, training duration, etc. table 2 presents the results of this comparison.

As we can see, the highest recognition rate is resulted when the number of neurons in the input layer is 18 and the number of neurons in the hidden layer is 10. Results of the neural network are presented in figure 5.

Table 2. Comparison of the results for different neural network training conditions

The average personality recognition rate in comparison to MMPI test		Training duration (seconds)	Number of trainings	Number of neurons in the hidden layer	Number of neurons in the input layer
Test	**Training**				
46%	69%	500	70	5	10
59%	72%	500	70	7	15
61%	76%	700	70	10	18
52%	74%	1100	70	15	20
57%	71%	2000	70	19	25

As we can see, the highest recognition rate is resulted when the number of neurons in the input layer is 18 and the number of neurons in the hidden layer is 10. Results of the neural network are presented in figure 5.

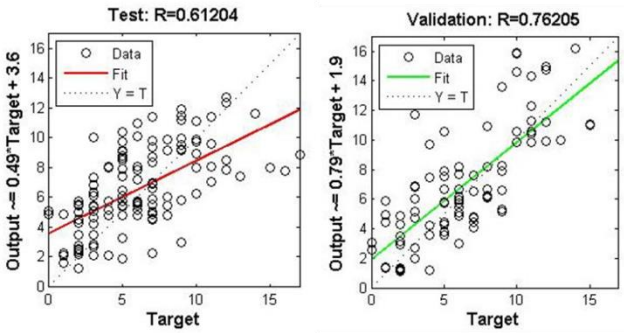


Figure 5. (a) Regression of the validation data (b) regression of test data

Figure (5a) presents the regression of validation data and figure (5b) presents the regression of the test data. Moreover, table 3 is presented to show the role of each feature extraction stage in improving the final output of the personality recognition system.

In this table, the input feature vector of the personality recognition system only consists of one feature extraction method mentioned in each row. The same feature vector is used to estimate the clinical and validity scales of the neural network. Therefore, according to the recognition rate of the personality recognition system in each row, we can find the effect of the feature extraction method on estimating the clinical and validity scales parameters.

According to table 3, we can find some important points. The sum of the values in each row of table 3 and normalizing it, shows the effect of each feature extraction stage on the final output of the personality recognition system.

As we can see, the HLAC feature has the highest effect (29.08%) and the ratio of vertical to horizontal words has the lowest effect (8.56%) on the accuracy of the propose system`s output. Moreover, the sum of the values of each column in table 3, whose results are presented in figure 6, shows the average estimated scales for each test of the proposed personality recognition system.

Table 3: The role of each feature extraction stage in improving the final output of the personality recognition system

Feature / Scale	*Ratio of vertical to horizontal*	*Character sizes*	*Word tilt*	*Line spaces*	*Word spaces*	*Margin value from top of page*	*Word extension*	*Line tilt*	*HLAC feature*
L	5%	6.5%	6.8%	8.5%	4.5%	3.1%	6.4%	6.5%	17.3%
K	4.5%	5.8%	6.2%	7.6%	5.1%	3.8%	5.6%	6.7%	16.8%
F	4.8%	6.1%	5.6%	6.4%	3.1%	4.2%	6.1%	5.9%	16.2%
Ma	5.1%	5.4%	5.9%	6.8%	4.7%	4.5%	4.9%	6.1%	17.5%
Sc	5.7%	4.6%	6.1%	5.2%	3.4%	4.1%	5.7%	5.3%	16.5%
Pt	6.1%	6.8%	6.4%	7.5%	4.6%	3.2%	4.9%	5.6%	16.4%
Pa	4.7%	5%	5.2%	6.9%	4.5%	4.5%	5.2%	6%	16.4%
Pd	4.2%	5.3%	5.5%	7.2%	5.1%	4.3%	5.8%	6.5%	17.1%
Hy	5.8%	4.9%	6.4%	6.9%	5.7%	3.5%	4.5%	5.9%	17.6%
D	5%	5%	5%	5%	5.1%	4.9%	4.6%	5.6%	16.5%
Hs	5.7%	3.4%	5.9%	6.3%	4.4%	5.4%	5.4%	5.8%	17.4%

In this figure, the clinical scale Hs has the highest value (64.60%) and the clinical scale Pa has the lowest estimation value (56.6%).

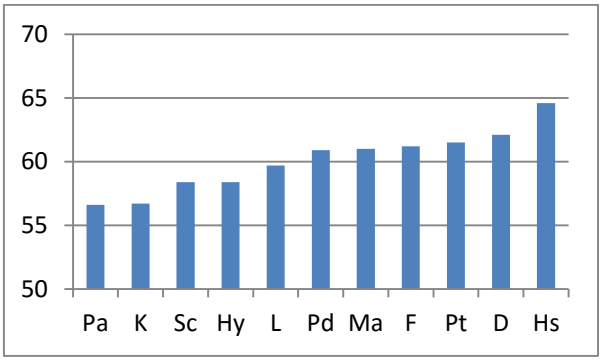


Figure 6. The average estimated scales for each test

Finally, the result of the proposed algorithm is compared with conventional methods using one database and table 4 presents the results.

Table 4. Evaluation of the proposed algorithm

Test	Efficiency	measure
13%	20%	K-Means
46%	58%	Fuzzy C means
61%	76%	Proposed method

8. Conclusions

This research employs a personality recognition system which automatically extracts the character parameters from Persian handwritten texts. Well as due to the use of valid personality test (MMPI) in the training system, the extracted personality parameters on the test step was pretty standard, in result the proposed method has superior finality to other methods. The proposed system create feature vector using HLAC independent feature, context features such as value of margin from the top, word extraction, character sizes, line space, word space, word tilts, horizontal to vertical ratio of characters, and lie tilts. The MLP neural network is used for classification; Such that the output of this network will be the parameters of the author characters. Training and evaluation prepared for the database system is used by 70 different writers.

considering the results, the advantage of the proposed algorithm is as follows:

1- Using dependent and Independent features of text in the process of feature extraction.

2- The proposed personality recognition system is automated, particularly in the process of feature extraction (None automated systems are very unendurable and time-consuming)

3- Enhance accuracy and reliability of the personality recognition system due to the use of MMPI test on training step

4- No need to segmentation on feature extraction phase

5- Using GDA to increase the space between classes

References

[1] S. Hock, K. Siang Teh , L.Y. Yee, “An Overview on the Use of Graphology as a Tool for Career Guidance”, CMU. Journal 2005, Vol. 4(1).

[2] A.B.Sharif, E. Kabir, “Computer Aided Graphology For Farsi Handwriting ”, Computer and electronic engineering journal (Persian), 2005 , Vol 3 , Num 2; Page(s) 73-79.

[3] Sh. Prasad, V. Kumar, A. Sapre, “Handwriting Analysis based on Segmentation Method for Prediction of Human Personality using Support Vector Machine”, International Journal of Computer Applications (0975 – 8887) , October 2010,Vol 8– No.12.

[4] J. Fisher, A. Maredia, A. Nixon, N. Williams, J. Leet, “Identifying Personality Traits, and Especially Traits Resulting in Violent Behavior through Automatic Handwriting Analysis”, Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 4th, 2012.

[5] A. Rahiman, D. Varghese , M. Kumar,” Handwritten Analysis Based Individualistic Traits Prediction”, International Journal of Image Processing (IJIP),2013, Volume (7) : Issue (2) , 2013.

[6] H.N. Champa , K.R.AnandaKumar , “Artificial Neural Network for Human Behavior Prediction through Handwriting Analysis”, International Journal of Computer Applications (0975 – 8887), May 2010, Vol 2 – No.2.

[7] Y.Norouz Zade, H.NezamAbadi,"Handwriting Graphology using image processing and fuzzy inference system",8th Intelligent Systems Conference, Ferdowsi University of Mashhad, Iran. 2007.

[8] I. A. Jannoud, "Automatic Arabic Hand Written Text Recognition System," *American Journal of Applied Sciences*, vol. 4, pp. 857-864, 2007.

[9] Tsai C-F, Wu J-W, (2007). "Using neural network ensembles for bankruptcy prediction and credit scoring. " Expert Systems with Applications,in press.164-173.

[10] T. Kurita, K. Hotta, and T. Mishima, "Scale and rotation invariant recognition method using higher-order local autocorrelation features of Log-Polar image," Proc. Third Asian Conference on Computer Vision, 1998, Vol 2, pp.89-96.

[11] S. M. Lajevardi, Zahir M. Hussain, "Facial Using GA to increase the distance between classes Expression Recognition: Gabor Filters versus Higher-Order Correlators," International Conference on Communication, Computer and Power (ICCCP'09), Muscat, Oman, 15-18 Feb. 2009.

[12] Arica Nazif ,Yamin-Vural Fatos T., "An overview of character recognition based focused on off-line handwriting", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, May 2001, Vol. 31, No. 2.

[13] Perssman, Lyons , Lavson & Strain, "Religions belief depression , and ambulation Status in elderly woman with broken hips". American Journal of psychiatry , 1990, 147, 758-66.

Behnam Fallah received his B.Sc. in computer engineering from department of computer Engineering, Hamedan Branch, Islamic Azad University, in 2008, and her M.Sc. in computer software engineering from department of computer, Qazvin Branch, Islamic Azad University, in 2015.His research interests includes Image processing, Wireless Networks and Data Mining.

Hassan Khotanlou is an associate professor in department of computer at Bu-Ali Sina University, Hamedan, Iran. He received his B.Sc. degree in computer engineering from IUST University in 1995, and his MSc. degrees in artificial intelligence engineering from Shiraz University in 1997 and his Ph.D. degrees in artificial intelligence engineering from Pierre & Marie Curie University (Paris VI – TELECO) in 2007.His main research interests are Image processing, Fuzzy Systems, Acoustic, Data Mining, Computer Networks, Medical Image processing and Artificial Intelligence.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

Select the most relevant input parameters using WEKA for models forecast Solar radiation based on Artificial Neural Networks

Somaieh Ayalvary[1], Zohreh Jahani[2] and Morteza Babazadeh[3]

[1] Department of Computer Engineering, ISLAMIC AZAD UNIVERSITY*Babol Branch*
Langerud*, 013/ *4253*, *Iran
Somaieh.Ayalvary@gmail.com

[2] Department of Computer Engineering, ISLAMIC AZAD UNIVERSITY*Babol Branch*
Gorgan*, 017/1235, *Iran
zohreh.jahanii@gmail.com

[3] Department of Computer Engineering, ISLAMIC AZAD UNIVERSITY*Babol Branch*
Babol*, 011/3241, *Iran
Morteza_babazade@yahoo.com

Abstract

Forecasting solar radiation is important for many applications in research related to renewable energy. Solar radiation is forecasted by solar radiation forecast models including the traditional models and artificial neural network (ANN) based model. There are geographical and meteorological variables that affect the solar radiation, thus identifying the appropriate variables to forecast solar radiation correctly is an important issue in the research area. Accordingly Waikato Environment for Knowledge Analysis (WEKA) Software was used in 11 points in Guilan based on different weather conditions to find the most effective input parameters to forecast solar radiation in different ANN models. Input parameters include latitude, longitude, maximum wind speed, average temperatures in each month, the average maximum air temperature, average minimum air temperature, sunshine, monthly rainfall, maximum rainfall in a day for different cities of Gilan. In order to check the reliability of the forecasts by known parameters, three ANN models have developed (ANN-1, ANN-2 and ANN-3). The maximum MAPE for ANN-1, ANN-2 and ANN-3 equals 22.15%, 20.29% and 22.14%, respectively indicating 1.86% improvement in the accuracy in the prediction of ANN-2.

Keywords: *Neural Network, Data Mining, WEKA*

1. Introduction

Solar energy is a clean source of energy with high potential to meet the needs related to energy. The assessment of solar potential energy requires information on solar radiation in different places. In order to find the most relevant parameters, the variables must be selected by combining various input parameters so that the best forecast is made which is time consuming. Therefore, in this study WEKA software version 7.3.10 is used to forecast solar radiation at 11 points in Guilan Province with different climatic conditions which is discussed below. In order to check the accuracy of prediction, ANN models have been developed (ANN-1, ANN-2 and ANN-3). Model ANN- 1 was created using input parameters while model ANN-2 was provided by the most relevant parameters presented by WEKA software and model ANN-3 was presented by removing the relevant parameters that can be used to forecast solar radiation at some pints in Guilan. The paper is set out as follows; Evaluation studies in section 2 and databases and methods are presented in Section 3. Results in Section 4 and conclude have come in Section 5.

2. Research about Identify input parameters to forecast Solar radiation based on ANN

ANN models use different meteorological and geographical variables of an area as the input data to forecast solar radiation. Azeez [1] used back propagation neural network to estimate average monthly solar radiation in Gusau, Nigeria. The duration of radiation, the average prevailing temperature and relative humidity were considered as output. Statistical analysis (R = 99.96, MPE = 0.8512, RMSE = 0.0028) showed the best agreement between measured and estimated values of solar global radiation.

Linares-Rodriguez And colleagues [2] used From MLP model estimate solar radiation in Spain by using irradiation obtained by satellites. The input layer has 12 inputs (11 channels Meteosat and radiation of the sun on a clear day). RMSE is equal to 6.74%. The model works well in cloud and sunny weather conditions. According to studies, it was found that the accuracy of forecasting ANN model change by using geographic and meteorological variables as input parameters. To select the relevant input parameters, the researcher must use various combinations of input parameters to assess the accuracy of forecast models ANN. That It requires a lot of computational analysis. Therefore choosing the most relevant input parameters for the ANN models is and important research failure addressed in this study.

3. Methods

3.1. Solar radiation data source

11 selected locations in different climate zones in Gilan which were used to try and test the ANN models, came in Figure 3. Data from these stations Meteorological Organization in Gilan Province in the study for an average of four years, from 1387 to 1390, are presented in tables 11 and 12.

3.2. Select the input variables using WEKA

Choosing the input variables is the first step is to develop ANN models. The Input test data including temperature, maximum temperature, minimum temperature, altitude, hours of radiation, latitude and longitude for the solar radiation models are obtained by Table 1. In the process of selection of variables, the most relevant input variables should be evaluated to forecast solar radiation. To select the related input variables the feature evaluator and search method are selected as the result of which all variables are observed. The rank of each input variable as determined by WEKA to forecast solar radiation is presented in figures 3, 6 and 9.
The latitude variables have lowest rank. So, At Select the relevant input variables To calculate the accuracy of forecasting solar radiation, Latitude removed from the input vector X.And The accuracy of prediction was calculated by using ANN based on relevant input variable. After solving the problem of selection of variables, Three ANN (ie ANN-1, ANN-2 and ANN-3) were developed to calculate predictive accuracy. ANN-1 model uses variables of average monthly air temperature (T), the average minimum temperature (Tmin), the average maximum temperature (Tmax), the maximum wind speed (meters per second)) wind (m / s , sunshine (hours) (SH), the maximum daily rainfall (mm), rain (mm) day monthly rainfall (mm), rain (mm) month and latitude. ANN-2 model uses From The most relevant variables obtained from WEKA(Average air temperature in each month (T), the average minimum temperature (Tmin), the average maximum temperature (Tmax), the maximum wind speed (meters per second)) wind (m / s, sunshine (hours) (SH), the maximum daily rainfall (mm), rain (mm) day monthly rainfall (mm) rain (mm) month) and ANN-3 model uses variables, the average minimum temperature (Tmin), the average maximum temperature (Tmax), the maximum wind speed (meters per second)) wind (m / s, sunshine (hours) (SH), the maximum daily rainfall (mm), rain (mm) day and monthly rainfall (mm) rain (mm) month.[3,4]

3.3. Predictive models of solar radiation with selective input

ANN Models (ANN-1, ANN-2 and ANN-3) have been created by network fitness tools that are used to forecast.
The number of neurons in the hidden layer is evaluated by equation (1) [5,6] where H_n and S_n are the number of hidden layer neurons and sample data used in the ANN model, I_n and O_n also indicated input and output parameters.

$$Hn = (In + On) / 2 + \sqrt{Sn} \quad (1)$$

Sensitivity tests to validate the number of hidden layer neurons is performed by calculating the change in prediction error (MAPE) at the time of change in the number of neurons in the hidden layer as ± 5 of hidden layer neurons calculated by the equation (1). The analysis of neurons' sensitivity is done for ANN models; MAPE was obtained by Equation 2 and ANN structure by minimum MAPE is used to forecast solar radiation.

$$\mathbf{MAPE} = \mathbf{((1/n \sum_{i=1}^{n} |H/H0 -} \hat{H} / \hat{H}0| \mathbf{) \times 100) / \hat{H}/ \hat{H}0} \quad (2)$$

4. Discussion

In this system, we mean absolute percentage error (MAPE) considered as a parameter. To help ranking software "Weka" we find that At Model ANN1, latitude has least importance.

Table 1: meteorological data and geographic coordinates for the 11 cities in the Gilan ANN1

City	Wind	T	Tmax	Tmin	SH
Astara	9.7	15.96	19.42	12.43	144.57
Anzali	16.67	17.08	19.41	14.7	149.84
Jirandeh	24.16	12.72	17.05	8.34	221.83
Rudesar	9.16	16.97	20.51	13.4	136.23
Rasht	12.16	16.74	21.3	12.12	130.85
Kiashahr	12.08	16.87	20.44	13.27	127.85
Talesh	11.58	16.59	19.84	13.09	116.34
Masuleh	10.66	12.55	16.19	8.86	111.22
Manjal	17.83	18.41	23.53	13.24	233.61
Lahijan	7.83	16.91	21.38	12.39	141.2
Dealaman	13.08	12.07	16.9	7.21	153.48

Table 2: meteorological data and geographic coordinates for the 11 cities in the Gilan ANN1

City	Rain month	Rain day	Lat	Long	H/H0	Ĥ/Ĥ0
Astara	94.27	41.44	55	38	0.5	0.6026
Anzali	123.97	38.47	49	37	0.41	0.6026
Jirandeh	21.83	7.79	50	37	0.56	0.6043
Rudesar	109.17	42.66	50	37	0.5	0.6021
Rasht	102.06	32.23	49	37	0.48	0.6019
Kiashahr	83.22	21.2	56	32	0.51	0.6018
Talesh	102.79	39.9	48	37	0.42	0.6012
Masuleh	75.75	20.29	49	37	0.43	0.6009
Manjal	13.33	6.46	49	36	0.51	0.6045
Lahijan	111.13	35.65	50	37	0.34	0.6023
Dealaman	33.04	14.14	49	37	0.5	0.6028

If the amount of solar radiation on the earth's surface (H / H0) is considered in the WEKA class software, the ranking is obtained as follows:
Where the longitude, latitude and average temperature are the input values that have the least impact on the results and have a lower rank.
Table 3 presents the input variable ranks by WEKA algorithm to predict solar radiation in the model ANN1 that contains all input values (Latitude, longitude, maximum wind speed, with average temperatures in each month, the average maximum air temperature, average minimum air temperature, sunshine hours, rainfall monthly, maximum rainfall in a single day as our workload to the system.)

Table 3: The number of input variables by WEKA algorithm to predict solar radiation in the ANN1

Rank	***attributes***
0.083	Wind
0.0495	Rain(mm)month
0.0324	SH
0.0188	Rain(mm)day
-0.0178	Tmin
-0.0217	Tmax
-0.0328	T
-0.0587	Lat
-0.0611	Long

In ANN2 model the longitude and latitude parameters must be removed, the maximum wind speed, average air temperature in each month, the mean maximum air temperature, the average minimum air temperature, hours of sunshine, monthly rainfall and maximum rainfall in one day are entered to the system as the workload.
For this purpose we perform data mining for 11 cities in Guilan province and remove two cities in which the input values for longitude and latitude are more important (in Manjil and Lahijan). In this model the temperature parameter has the least importance.

Table 4: meteorological data and geographic coordinates for the city of Gilan 9 ANN2 model

City	Wind	T	Tmax	Tmin
Astara	9.7	15.96	19.42	12.43
Anzali	16.67	17.08	19.41	14.7
Jirandeh	24.16	12.72	17.05	8.34
Rudesar	9.16	16.97	20.51	13.4
Rasht	12.16	16.74	21.3	12.12
Kiashahr	12.08	16.87	20.44	13.27
Talesh	11.58	16.59	19.84	13.09
Masuleh	10.66	12.55	16.19	8.86
Dealaman	13.08	12.07	16.9	7.21

Table 5: meteorological data and geographic coordinates for the city of Gilan 9 ANN2 model

City	SH	Rain month	Rain day	H/H0	Ĥ/Ĥ0
Astara	144.57	94.27	41.44	0.5	0.6026
Anzali	149.84	123.97	38.47	0.41	0.6026
Jirandeh	221.83	21.83	7.79	0.56	0.6043
Rudesar	136.23	109.17	42.66	0.5	0.6021
Rasht	130.85	102.06	32.23	0.48	0.6019
Kiashahr	127.85	83.22	21.2	0.51	0.6018
Talesh	116.34	102.79	39.9	0.42	0.6012
Masuleh	111.22	75.75	20.29	0.43	0.6009
Dealaman	153.48	33.04	14.14	0.5	0.6028

Table 6: The number of input variables by WEKA algorithm to predict solar radiation in the ANN2

Rank	***attributes***
0.083	Wind
0.0495	Rain(mm)month
0.0324	SH
0.0188	Rain(mm)day
-0.0178	Tmin
-0.0217	Tmax
-0.0328	T

Temperature parameter should be eliminated in model ANN3,The maximum wind speed, the mean maximum air temperature, hours of sunshine, monthly rainfall and maximum rainfall in one day are entered to the system as the workload.
For this purpose one of the 11 cities in Guilan province in which temperature parameter has the highest importance (Jirandeh) is eliminated.

Table 7: meteorological data and geographic coordinates for the 8 cities of Gilan in the ANN3

City	Wind	T	Tmax	Tmin	SH
Astara	9.7	15.96	19.42	12.43	144.57
Anzali	16.67	17.08	19.41	14.7	149.84
Rudesar	9.16	16.97	20.51	13.4	136.23
Rasht	12.16	16.74	21.3	12.12	130.85
Kiashahr	12.08	16.87	20.44	13.27	127.85
Talesh	11.58	16.59	19.84	13.09	116.34
Masuleh	10.66	12.55	16.19	8.86	111.22
Dealaman	13.08	12.07	16.9	7.21	153.48

Table 8: meteorological data and geographic coordinates for the 8 cities of Gilan in the ANN3

City	Rain month	Rain day	Lat	Long	H/H0	Ĥ/Ĥ0
Astara	94.27	41.44	55	38	0.5	0.6026
Anzali	123.97	38.47	49	37	0.41	0.6026
Rudesar	109.17	42.66	50	37	0.5	0.6021
Rasht	102.06	32.23	49	37	0.48	0.6019
Kiashahr	83.22	21.2	56	32	0.51	0.6018
Talesh	102.79	39.9	48	37	0.42	0.6012
Masuleh	75.75	20.29	49	37	0.43	0.6009
Dealaman	33.04	14.14	49	37	0.5	0.6028

Table 9: The number of input variables by WEKA algorithm to predict solar radiation in the ANN3

Rank	***attributes***
0.083	Wind
0.0495	Rain(mm)month
0.0324	SH
0.0188	Rain(mm)day
-0.0178	Tmin
-0.0217	Tmax

With the help of equation (2) to obtain the mean absolute percent error for three ANN1, ANN2 and ANN3 The mean absolute percent error for the ANN1:

$$MAPE = (1 / n \text{ Ɛ } ni = 1 \mid (H / H0 - \hat{H} / \hat{H}0) / \hat{H} / \hat{H}0 \mid) \times 100 = 22.15 \quad (3)$$

The mean absolute percent error for the ANN2:

$$MAPE = (1 / n \text{ Ɛ } ni = 1 \mid (H / H0 - \hat{H} / \hat{H}0) / \hat{H} / \hat{H}0 \mid) \times 100 = 20.29 \quad (4)$$

The mean absolute percent error for the ANN3:

$$MAPE = (1 / n \text{ Ɛ } ni = 1 \mid (H / H0 - \hat{H} / \hat{H}0) / \hat{H} / \hat{H}0 \mid) \times 100 = 22.14 \quad (5)$$

Model ANN2 is able to better forecast solar radiation for having the lowest error.
Based on the result of the calculation we find that the ANN2 model outperforms model ANN3 and the performance of model ANN3 is higher than ANN1.
The error results indicate that model ANN2has the lowest error and it is considered as the best model.
We want to obtain the maximum and minimum MAPE error obtained by equation (2) using algorithms GA and SA described below.
For this purpose the values of H / H0 and Ĥ / Ĥ0 are considered in the range between 0 and 1.
For each of these functions Simulated Annealing (SA) algorithms and genetic algorithm are used and compared.
This algorithm is made of two nested loops. First the temperature is very high (in our simulation randi (100,000)), so particle displacement is very high.
With decreasing temperature (Increasing the number of reps and getting closer to the answer), particle displacement is reduced and local search will occur.
One major difference between SA and GA is that SA is a single factor algorithm while GA is a population or multi-factor algorithm, thus GA gets a better answer and the possibility of local minimum is low. For example, in our simulations, SA algorithm is repeated more than 15,000 times in each run to reach the optimal solution, while the genetic algorithm is run 10 times and assuming a total of 40 children in each run, it is performed 400 times and has reached the optimal solution[7].

Simulation equation 2 with SA

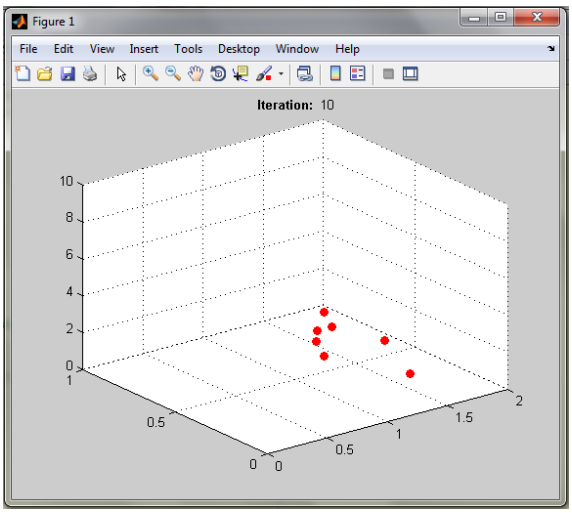


Figure 1: Output Simulation shows the minimum MAPE error.

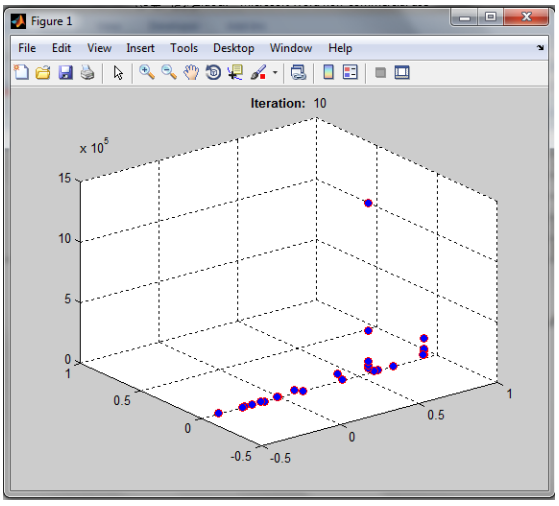


Figure 2: Output Simulation shows the maximum MAPE error

Table 10: Abbreviations used

Lat	Latitude
Wind	Longitude
T	The average air temperature per month
Tmax	The average maximum air temperature
Tmin	The average minimum air temperature
SH	Sunshine hours(hrs)
Rain month	Monthly Rainfall(mm)
Rain day	The maximum daily rainfall(mm)
MAPE	The mean absolute percentage error
H/H0	Monthly average daily solar radiation forecast data for the month i
$\hat{H}/\hat{H}0$	Monthly average daily solar radiation data measured for month i

Figure 3: The above image is a map of Gilan That shows the selected cities to examine and test the model ANN.

Table 11: meteorological data and geographic coordinates for the 11 cities in Gilan

City	Wind	T	Tmax	Tmin	SH
Astara	9.7	15.96	19.42	12.43	144.57
Anzali	16.67	17.08	19.41	14.7	149.84
Jirandeh	24.16	12.72	17.05	8.34	221.83
Rudesar	9.16	16.97	20.51	13.4	136.23
Rasht	12.16	16.74	21.3	12.12	130.85
Kiashahr	12.08	16.87	20.44	13.27	127.85
Talesh	11.58	16.59	19.84	13.09	116.34
Masuleh	10.66	12.55	16.19	8.86	111.22
Manjal	17.83	18.41	23.53	13.24	233.61
Lahijan	7.83	16.91	21.38	12.39	141.2
Dealaman	13.08	12.07	16.9	7.21	153.48

Table 12: meteorological data and geographic coordinates for the 11 cities in Gilan

City	Rain month	Rain day	Lat	Long	H/H0
Astara	94.27	41.44	55	38	0.5
Anzali	123.97	38.47	49	37	0.41
Jirandeh	21.83	7.79	50	37	0.56
Rudesar	109.17	42.66	50	37	0.5
Rasht	102.06	32.23	49	37	0.48
Kiashahr	83.22	21.2	56	32	0.51
Talesh	102.79	39.9	48	37	0.42
Masuleh	75.75	20.29	49	37	0.43
Manjal	13.33	6.46	49	36	0.51
Lahijan	111.13	35.65	50	37	0.34
Dealaman	33.04	14.14	49	37	0.5

The first model based on the parameters of the sundial, estimates the amount of radiation on a horizontal surface is Angstrom empirical equation (2) and Prescott (8).

$$\hat{H} / \hat{H}0 = a + b\,(\tilde{n} / \tilde{N}) \qquad (6)$$

In the above equation $\hat{H}$ represents the total daily radiation per month, $\hat{H}0$ represents measured radiation outside the atmosphere, $\tilde{N}$ is the average monthly hours of sunshine daily, $\tilde{N}$ is the average monthly peak sunshine hours (during the day).[8]

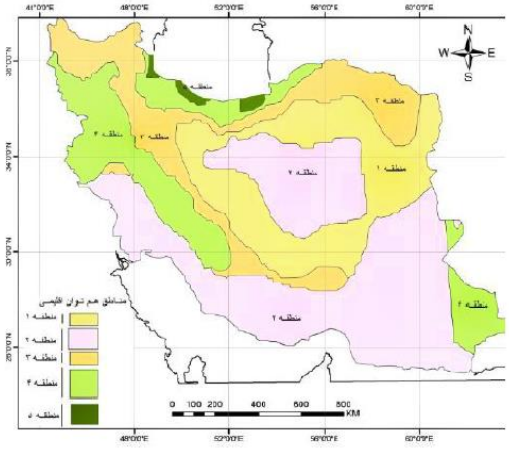


Figure 4: climatic classification

In the above map Iran is divided into five regional areas Gilan province is in the region of five. The parameters a and b are coefficients fixed equation That at Region Five are Equal 0.404 and 0.204 respectively.

$$H / H0 = (KT) \times (TD)^{0.5} \qquad (7)$$

$$TD = Tmax - Tmin \qquad (8)$$

In the above equation are H and H0 radiation reaching the Earth and Extraterrestrial radiation in calories per square centimeter per day (Cal \ cm2 day), respectively. TD is Daily temperature range (0C) and KT is constant coefficient equation that

For coastal and non-coastal are 0.19 and 0.16 respectively.

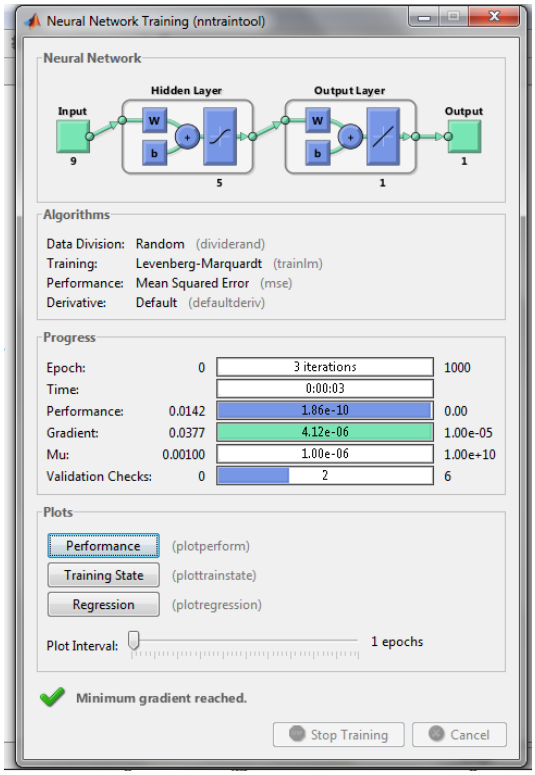


Figure 5: Evaluation of neural networks ANN-1

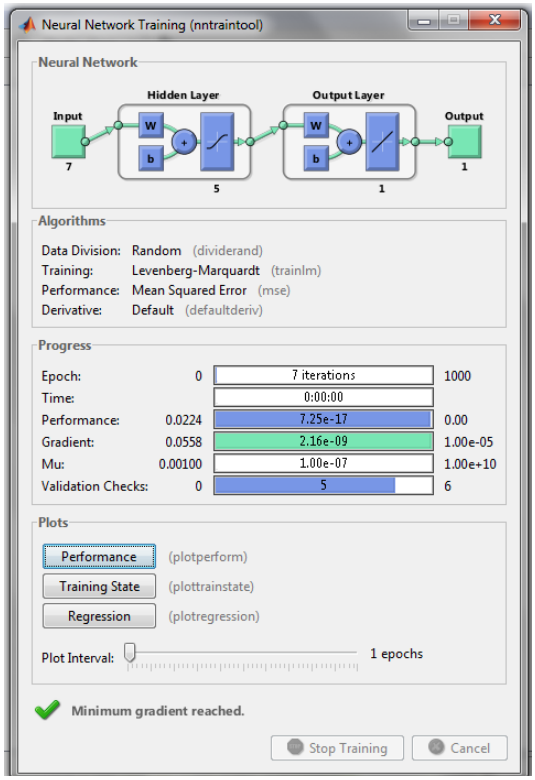


Figure 6: Evaluation of neural networks ANN-2

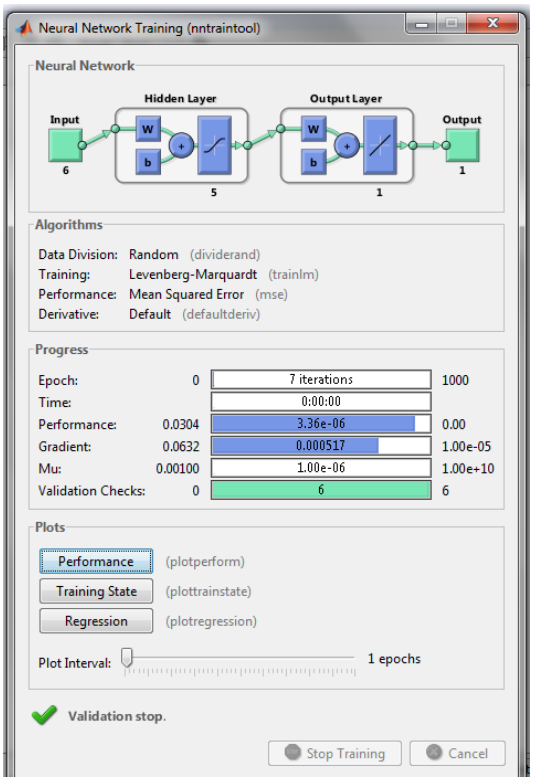


Figure 7: Evaluation of neural networks ANN-3

5. Conclusion

The present study shows the robust nature of WEKA in evaluating the most effective parameter to forecast solar radiation using ANN. It was found that the most relevant parameters to forecast solar radiation include temperature, maximum temperature, minimum temperature, altitude, and the hours of sunshine.Maximum MAPE for models ANN-1, ANN-2 and ANN-3 equals 22.15%, 20.29% and 22.14%, respectively indicating a high level of accuracy for ANN-2 that have used the most relevant input variables. ANN-2 model developed could be used to forecast solar radiation at any location in Guilan. Further studies can be made for more accurate estimation of the solar potential of the area. The future study should be focused on finding the most relevant parameters of the meteorological variables with improved forecast accuracy in different ANN models. We made a 5 layer neural network by data collected from 11 cities in Guilan. Input reduction and elimination of longitude and latitude from the input data caused the model ANN2 to be built more quickly than ANN1. As a result in this model we observe higher performance ability (Throughput).[9]

References

[1] Azeez MAA. Artificial neural network estimation of global solar radiation using meteorological parameters in Gusau, Nigeria. Arch Appl Sci Res 2011;3 (2):586–95.

[2] Linares-Rodriguez A, Ruiz-Arias JA, Pozo-Vazquez D, Tovar-Pescador J. An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images. Energy 2013;61:636–45.

[3] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update; SIGKDD Explorations 2009; 11(1). [accessed 02.02.2013] ⟨http://dx.doi.org/http://www.cs.waikato.ac.nz/~eibe/ pubs/weka_update.pdf⟩.

[4] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. A Book of Morgan Kaufmann Publishers. 3rd ed.2011.

[5] Chow SKH, Lee EWM, Li DHW. Short-term prediction of photovoltaic energy generation by intelligent approach. Energy Build 2012;55:660–7.

[6] Frederick M. Neuroshell 2 Manual, Ward Systems Group Inc., 1996.

[7] Bakirci K. Models of solar radiation with hours of bright sunshine: a review. Renew Sustain Energy Rev 2009;13:2580–8.

[8] Khatib T, Mohamed A, Sopian K. A review of solar energy modeling techni-ques. Renew Sustain Energy Rev 2012;16:2864–9.

[9] Rehman S, Mohandes M. Estimation of diffuse fraction of global solar radiation using artificial neural networks. Energy Sources, Part A 2009;31: 974–84.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

New Method Of Feature Selection For Persian Text Mining Based On Evolutionary Algorithms

Akram Roshdi
Department of Computer, Islamic Azad University, Khoy Branch,Iran
Akram.roshdi5@gmail.com

Abstract

Today, with the increasingly growing volume of text information, text classification methods seem to be essential. Also, increase in the volume of Persian text resources adds to the importance of this issue. However, classification works which have been especially done in Persian are not still as extensive as those of Latin, Chinese, etc. In this paper, a system for Persian text classification is presented. This system is able to improve the standards of accuracy, retrieval and total efficiency. To achieve this goal, in this system, after texts preprocessing and feature extraction, a new improved method of feature selection based on Particle Swarm Optimization algorithm (PSO) is innovated for reducing dimension of feature vector. Eventually, the classification methods are applied in the reduced feature vector. To evaluate feature selection methods in the proposed classification system, classifiers of support vector machine (SVM), Naive Bayes, K nearest neighbor (KNN) and Decision Tree are employed. Results of the tests obtained from the implementation of the proposed system on a set of Hamshahri texts indicated its improved precision, recall, and overall efficiency. Also, SVM classification method had better performance in this paper.

Keywords: *Feature vector, classification, support vector machines, Feature Extraction, Dimensions Reduction.*

1. Introduction

In text classification, before using any method, it is important to convert texts into a suitable display form. After such a conversion, feature selection algorithms are applied and texts are classified using the selected features. In this study, an improved model based on Particle Swarm Optimization algorithm is used to select the text feature. In fact, in this study, the mutation operator (based on genetic algorithm) is employed to improve the speed and accuracy of convergence of Particle Swarm Optimization algorithm. Methods and strategies used in the course of this investigation are given in the following sections.

2. Previous works

Due to complex structural problems of Persian language, fewer studies have been undertaken in the field of text mining in Persian than other languages. Classifying Persian texts is one of the areas affected by these problems and few works have been done in this regard so that, compared to the previous works in other languages, this issue seems to be a rather new research area. In the field of text classification, previous studies are, essentially, based on two-class classification technique. Some training methods in accordance to two-class strategy are Naive Bayes, K nearest neighbor (KNN), support vector machine (SVM), Decision Tree, and etc. Most researches on text classification, were designed, carried out and tested on English articles. A number of training techniques for text classification have been also implemented on other European languages such as German, Italian and Spanish. Some other techniques have been conducted on Asian languages such as Chinese and Japanese. In confrontation with complex structures and large data volume, statistical methods do not work. Since this issue has a number of features, some feature selection methods must be used which can reduce the number of features. Therefore, feature selection is one of the most important steps in text classification. For feature selection, several methods exist. Recently, researchers pay much attention to the evolutionary algorithms in the feature selection. Evolutionary algorithms can be used for image processing tasks. For example, genetic algorithm is used for face recognition system. Naive Bayes and K nearest neighbor (KNN), along with genetic algorithm, are used for to remove difficult-to-learn data. Three combinatory methods have been investigated on five Standard English data. The results indicate the non-relevant data elimination. In all these papers, it has been shown that the efficiency in evolutionary algorithms for feature selection is more than traditional statistical methods. Persian language due to its complex structures, in accordance to other languages, has few studies. Classification of Persian texts is among the fields which are affected by this problem. A list of studies on the Persian language is presented in [1, 2]. An automated technique for the detection of stop words in Persian language is presented [3]. In the paper [4], a new method for root detection with a bottom-up strategy is presented for Persian language. Test results show that this method

has good flexibility. In Persian documents, statistical information of documents is used to retrieve and rank the documents. The efficiency of various information retrieval models, such as vector space model, possible space model, the language model [5], fuzzy model [5], is evaluated on a set of Persian documents.

3. Classifying text documents

Text classification includes several steps: preprocessing phase prepares documents for classification process that is similar to many natural language processing issues, in which irrelevant tags and words are removed. Indexing phase uses a weighting scheme to weigh statements in the text. The following classification step is selecting appropriate feature space among the statements in documents, which is a vital step; also, system precision largely depends on selection keys that show the document.
In the next classification step, documents are divided into training and testing parts; the former is used to train the system to recognize different patterns of classes and the latter is used for system evaluation. Classification process depends upon the applied algorithms [6]. Figure 1 shows a general overview of classification steps.
classification steps.

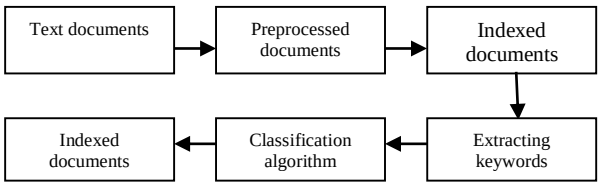


Fig.1 An overview of classification steps

4. Feature extraction methods

To apply text classification and feature extraction methods, an appropriate structure must be selected for displaying documents. The simplest and most conventional document display method is creating a feature space using all the words of a document. In this space of features, after removing special words and sometimes doing the etymology, a list of all words is made in the text and every document is displayed in different ways based on the words existing on the list and the weight of these words in the document. Table 1 shows text document display method in the form of vectors [7]. As seen in this table, the set { F1,…,Fn} indicates space of features and F_i shows a word which has been used at least once in the text. Moreover, the set { D1,…,Ds } shows a set of documents displayed in this vector display method. Each Wij value shows the weight of word Fj in document Di. In fact, there are different criteria for word weighing.

Table 1: Display method of vectors of text documents

words space				
$\mathbf{F_N}$	-	F_2	F_1	
$\mathbf{W_{1N}}$	-	W_{12}	W_{11}	D_1
$\mathbf{W_{2N}}$	-	W_{22}	W_{21}	D_2
-	-	-	-	-
$\mathbf{W_{SN}}$	-	W_{S2}	W_{S1}	D_S

For example, TF weighting scheme studies the number of feature repetition in a document, while TF-IDF studies feature repetition rate in other documents to determine its rate in the document. TFC also involves document length in its feature weight. All of these criteria try to determine the information importance of each feature in each document so that the classifier can classify documents in different classes based on the similarities of these scores. Texts cannot be directly evaluated by classifiers and are thus converted into appropriate display forms. As shown in Equation 1, text dj is usually displayed as a vector of word weight [8]:

$$d_j = \{ w_{1j}, w_{2j}, \ldots, w_{|T|j} \} \quad (1)$$

In Equation 1, T is a set of words (features) that is repeated at least once in at least one text of the training set and $0 \le wij \le 1$ indicates the weight allocated to the i^{th} word in the j^{th} text. Word weights are obtained using TFIDF normalized function. This method was first proposed in data recovery. Then, they were used in document classification for weighting features [9].
TFIDF method is one of the most conventional feature weighting methods which is obtained from the combination of TF- and IDF-based methods as follows [9].

$$wkj = \frac{tfidf(t_k d_j)}{\sqrt{\sum_{s=1}^{|T|}\left(tfidf(t_s d_j)\right)^2}} \quad (2)$$

where

$$tfidf(t_k, d_j) = \#(t_k, d_j).Log\frac{|Tr|}{\#Tr(t_k)} \quad (3)$$

In Equation 3, $|T_r|$ is the number of texts in the training set; the number of times word t_k is repeated in text d_j is # (t_k, d_j). T_r is training set and $|T_r|$ is its length. Also, # Tr (t_k) is the number of training set texts, in which word t_k occurs [8].

5. Evaluation criteria

In general, feature selection methods are applied as a step before the learning of classifiers. Effect of feature

selection methods and document display methods is obtained using efficiency measurement of each of the classifiers. In order to measure efficiency, standard definitions of precision (P), recall (R), and Fβ function are used[10].

$$P = \frac{\text{number of correctly found classes}}{\text{total number of found classes}} \quad (4)$$

$$R = \frac{\text{number of correctly found classes}}{\text{total number of correct classes}} \quad (5)$$

$$F\beta = \frac{(\beta^2+1)*P*R}{\beta^2*P+R} \quad (6)$$

P shows precision of classifications and R shows completion rate of the found set. High value of both of these factors indicates a high level of classification method; however, higher precision value is usually accompanied by reduced recall rate and high recall rate usually involves reduced precision. Depending on the application, sometimes high precision is important and sometimes higher recall is more optimal. Therefore, to attribute different weights to precision and recall, Fβ criterion can be used; considering the application and importance of each of these two factors (precision and recall), different weights can be assigned to them. In many academic researches, F1 criterion is used that gives equal weights to both of these factors. In this paper, this criterion was used for various evaluations.

6. Materials and methods

In this article, second version of Hamshahri's standard dataset was used. This body of text included more than 318 pieces of news between 1996 and 2007. The proposed algorithm is an improved model based on Particle Swarm Optimization algorithm for feature selection in text. In fact, in this paper, to improve the convergence speed and accuracy of Particle Swarm Optimization algorithm, the mutation operator (in accordance to genetic algorithms) is used. The proposed method has been named PM (Proposed Method). The proposed method has the appropriate global search (moving particles using particle swarm algorithm relations) and local search (due to the mutation applied on particles) at the same time. By adding the mutation operator to the Particle Swarm Optimization algorithm, the likelihood of escaping from local minimum increased greatly and this leads to likelihood enhancement of achieving an optimal solution. For increased accuracy, retrieval and overall efficiency of classification algorithms, several pre-processing methods such as document indexing and extra words deletion are used in training course construction. For feature selection, three methods of genetic algorithm and particle swarm optimization algorithm and the proposed method are used. Finally, 4 methods of classification as SVM, Naive Bayes, KNN and Decision Tree are used for document classification. Also, for preprocessing and classification stages, C# and MATLAB software were used, respectively. Afterward, precision, recall, and overall efficiency of classification algorithms were evaluated using the testing set.
Figure 2 shows the framework of the proposed classification system.

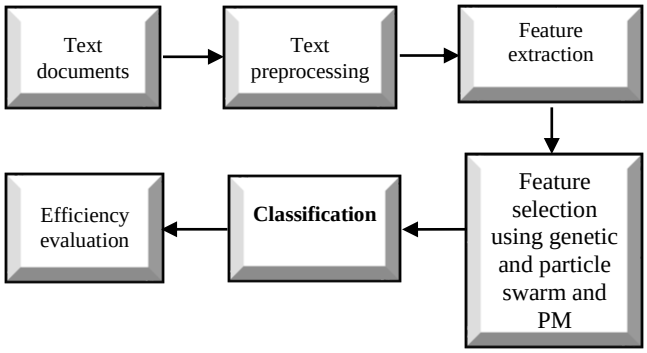


Fig.2 Framework of the proposed classification system

7. Results and discussion

Table 2, 3, 4 and 5, respectively, the comparison between efficiency, precision, retrieval and overall efficiency and SVM, Naive Bayes, KNN and Decision Tree are given.

Table2: Comparing efficiency (precision, recall, and F1 measurement) using SVM

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	76.54	76.54	76.54	76.40	74.68	75.53	77.75	76.58	77.16
10	77.24	75.95	76.59	75.25	74.05	74.65	78.83	76.58	77.69
15	77.92	77.85	77.88	78.18	77.22	77.69	79.11	79.28	79.19
20	78.49	76.58	77.52	78.42	77.85	78.13	80.01	80.38	80.19
25	76.54	75.95	76.24	79.09	77.85	78.46	80.92	79.75	80.33
30	79.50	74.68	77.01	82.13	81.65	81.89	82.28	82.28	82.28
35	81.44	80.38	80.91	81.21	81.01	81.11	83.83	83.54	83.68
40	81.26	79.75	80.50	83.72	83.54	83.63	84.63	82.28	83.44
45	81.02	79.75	80.38	83.77	83.54	83.66	84.01	84.38	84.19

Table3: Comparing (precision, recall, and F1 measurement) using naive Bayesian

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	72.14	70.25	71.19	70.73	70.26	70.49	69.74	69.62	69.68
10	72.40	67.72	69.98	73.03	70.25	71.61	73.79	71.52	72.64
15	75.65	71.52	73.53	73.18	71.52	72.34	75.88	74.05	74.95
20	72.04	68.35	70.15	73.73	72.78	73.25	74.47	72.78	73.62
25	74.55	72.78	73.66	74.75	72.15	73.43	75.51	71.52	73.98
30	75.04	69.62	72.23	74.83	72.15	73.47	77.62	73.95	75.74
35	76.17	71.52	73.77	76.32	72.15	74.18	78.98	75.29	77.09
40	77.51	73.42	75.41	76.21	72.78	74.45	77.99	76.01	76.99
45	76.34	68.99	72.48	75.17	71.52	73.77	77.52	74.60	76.03

Table4: Comparing efficiency (precision, recall, and F1 measurement) using KNN

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	70.69	69.62	70.15	75.64	74.05	74.84	77.79	75.95	76.86
10	75.69	72.15	73.88	76.50	75.32	75.90	78.83	76.58	77.69
15	76.67	75.32	75.99	78.32	77.22	77.76	79.37	78.48	78.92
20	75.64	74.05	74.84	80.03	79.11	79.57	79.99	77.85	78.90
25	78.33	77.22	77.77	79.04	77.22	78.12	80.92	79.75	80.33
30	79.86	77.85	78.84	80.43	79.11	79.76	82.28	82.64	82.28
35	81.25	79.78	80.49	81.13	80.38	80.76	82.60	81.75	82.17
40	81.97	79.75	80.84	81.92	80.38	81.14	83.65	83.11	83.38
45	79.80	79.11	79.46	81.01	81.84	82.10	82.94	82.28	82.61

Table5: Comparing (precision, recall, and F1 measurement) using decision tree classifier

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	69.45	68.99	69.22	74.19	72.78	73.48	75.98	74.05	74.51
10	72.18	67.72	69.88	76.77	74.68	75.71	77.02	75.95	76.48
15	73.63	66.46	69.86	77.77	74.05	75.86	78.41	75.32	76.83
20	74.55	72.78	73.66	75.10	73.42	74.25	76.56	75.95	76.26
25	75.82	74.68	75.25	78.23	75.32	76.74	78.41	75.32	76.83
30	75.13	71.52	73.28	78.83	76.58	77.69	79.01	78.48	78.74
35	76.71	76.58	76.65	79.62	74.05	76.74	80.03	79.11	79.57
40	74.94	74.68	74.81	78.33	77.22	77.77	78.46	73.42	75.85
45	74.12	74.12	74.12	78.08	72.15	75.00	78.60	74.68	76.59

According to the obtained results, it can be seen that SVM classifier had better performance than all other classifiers, as indicated by the tables given in the previous section. After SVM, KNN classifier had better performance than the two others. Figure 3 also shows the same issue.

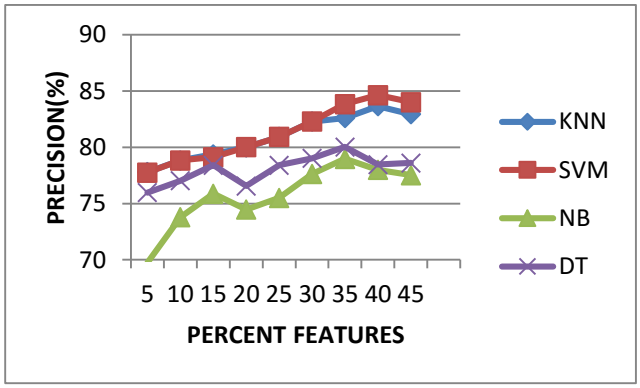


Fig.3 Comparing precision SVM ,KNN,DT and NB With proposed method

8. Conclusions

In this paper, a brief investigation was done about feature extraction methods and types of classification methods. All experiments were conducted on Persian documents of Hamshahri (citizen) standard data. Experimental results show that the proposed method has a good performance

and the precision, retrieval and efficiency of support vector machine classifier enjoy better function.
In future work, to improve the precision, retrieval and overall efficiency, we intend to generalize the proposed method for text categorization with more classes (for example 10 or 20 classes) using more features extraction. This is certainly a great step in increasing the accuracy of the researches in the field of information retrieval in Persian language. In the following, by examining other evolutionary algorithms and comparing them with the proposed algorithm for feature selection and providing a combinatory method for feature selection, we intend to increase the accuracy and efficiency of classification algorithms.

References

[1]. M .Aci, And , M . Avci," A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm",Elsevier, 2010,vol. 37, p.5061–5067.
[2]. M. shamsfard ,"processing Persian text: past finding and future challenges", Tehran universitypress, 2007.
[3]. A.yoosofan and M. zolghadri,"an automatic method for stopword recognition in Persian language", amirkabir university press, 2005.
[4]. M.Aljaly and O.frieder," improving the retrieval effectiveness via light stemming approach", journal of information science,2004,vol. 158, pp. 69-88.
[5]. A.Unler and A. MuratA," maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification",Elsevier,2011, No. 181, P. 4625–4641.
[6] M.saleh," list of dissertations on Persian language and computers", Tehran university press, 2007.
[7] M. shamsfard ,"processing Persian text: past finding and future challenges", Tehran university press,2007.
[8]. M. Litvakand and. M. last ,"classification of web documents using concept extraction from ontologies",Proceedings of the 2nd international conference on Autonomous intelligent systems: agents and data mining, Russia, , 2007, pp. 287-292.
[9]. V. Gupta, and S. lehal," a survey of text mining technique and applications",journal of emerging technologies in web intelligence,2009,vol.1, no.1.
[10]. A. Sharma , Sh. Dey," Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", International Journal of Computer Applications on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012.

Akram Roshdi received his B.Sc. in computer engineering from Shabestar University of Applied Science And Technology, Tabriz, Iran, in 2003, and his M.Sc. in Computer Engineering from Islamic Azad University, Shabestar branch, Tabriz, Iran, in 2013. She is Currently a PhD student at Department of Engineering, Qom branch, Islamic Azad University. His research interests include wireless networks, cloud computing and Data mining.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A new algorithm to create a profile for users of web site benefiting from web usage mining

Masomeh khabazfazli[1,2], Dr. Ali Harounabadi[3] and Dr. Shahram Jamali[4]

[1] Department of computer, Ardabil Science and Research branch, Islamic Azad University, Ardabil, Iran
[2] Department of computer, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Ma.fazli64@gmail.com
[3] Department of Computer Science, Islamic Azad University, Central Tehran Branch, Tehran, Iran
A.harounabadi@gmail.com
[4] Computer Engineering Department, University of Mohaghegh Ardabil, Ardabil, Iran
Jamali@iust.ac.ir

Abstract

Upon integration of internet and its various applications and increase of internet pages, access to information in search engines becomes difficult. To solve this problem, web page recommendation systems are used. In this paper, recommender engine are improved and web usage mining methods are used for this purpose. In recommendation system, clustering was used for classification of users' behavior. In fact, we implemented usage mining operation on the data related to each user for making its movement pattern. Then, web pages were recommended using neural network and markov model. So, performance of recommendation engine was improved using user's movement patterns and clustering and neural network and Markov model, and obtained better results than other methods. To predict the data recovery quality on web, two factors including accuracy and coverage were used

Keywords: *Web Page Recommendation, Web Mining, Web Usage Mining, Clustering, Neural Network, markov model*

1. Introduction

Upon establishment and expansion of internet and subsequently websites enhancement and upraise of its applications by users, searching the contents among extensive information on internet pages has become difficult for web users. The users face a great volume of recovered data. On the other side, topics such as purposeful advertisements and awareness of users' information are very important. Therefore, web mining is propounded for solving this problem. Web mining is the process of unknown and useful knowledge discovery through web data. Currently, broad researches have been applied in this relation and their purpose is solving problems related to data recovery [1].

One of objectives of this study is web pages recommendation for users and time and cost saving as well as better support of purposeful advertisements and electronic business. Thus, upon using web mining methods, this group of problems are solved somewhat. Users may select pages recommended to them which are related to the subject. In this part, generalities, objectives and necessities of this research w analyzed. In second part, background of study is explained. Third part includes main idea for offering web recommendation engine based on web usage mining. In fourth part, evaluation and results of experiments are provided and compared to other methods, and in final part summary of paper is provided.

2. RELATED WORK

Recently, web usage mining techniques are used extensively for prediction of internet pages. Access patterns are discovered from record file using methods such as association rules mining, clustering etc. that is used for prediction of users' behavior. In [2], a web usage mining method was used offered therein clustering was used so that users' behavior was clustered based on measurement set of log file data similarity to be used for prediction of internet pages. In fact, clustering has been made using similarity of above approximation. In this process, clustering was provided based on subject and shows common interests in each cluster. In [3], web usage mining techniques were used and analyzed the problems from two aspects including improvement of search engine through static saving of search results and weblog posts. This study offered search coverage method and used graph for recommendation to users. In [4], rough set theory was used for log file processing and keywords, upon combining two methods of content and collaborative filtering based recommender systems through design of two-layer graph that was made along with graph partitioning. Each node in pages later and users' layer respectively shows web pages and users. Therefore, similarity between pages and users is obtained by this way in graph partitioning.

3. Background

In this part, background contents which are important for understanding the offered method are raised. At first, web mining, then personalization based on usage mining and at end clustering approaches and neural network and markov model will be explained.

3.1 Web mining

Web mining is a subset of data mining technique for covering the web patterns that based on web pages analysis and which part of web data is explored is divided in three parts including web structure mining, web content mining and web usage mining [5].
Web content mining is discovery and extraction of useful data from web pages. Web structure mining discovers and analyzes the model. In this kind of web mining, web is modeled as a graph therein web pages are assumed as graph nodes and links between pages as graph edges [5]. Web usage mining includes discovery of user's access patterns for web server recording file.

3.2 Personalization based on web usage mining

Personalization is as one of the application fields of web exploring by which pages contents can be changes in accordance with users' interests in order to provide the internet services in a better way and also to meet the needs of users quickly. Several web personalization systems have been created based on web mining that all of them include two main stages: in the first stage which is performed offline, training data taken from user behavior on the Web are explored in order to detect the access patterns and extract the users' model. In the second stage which is done online, the model extracted from the first stage is used for interpretation and comparison with the traversal pattern of active user and then propositions are provided based on this comparison. The purpose of web personalization based on exploring the web application is offering a set of objects to the current user with orientation towards the user's preferences and interests.

3.3 Clustering

Clustering or cluster analysis is the process of grouping some physical and virtual objects in classes of similar objects. A cluster includes a set of data objects which are similar to each other. In general, two types of clustering (transaction clustering and page visiting clustering) can be applied for the transaction data of web application.

Each of these approaches has different applications and in particular, both of the two approaches can be used for web personalization. K-Means algorithm is one of the most important clustering algorithms that are widely used. In this algorithm, the samples are divided into k clusters and the number of k has already been specified.

3.4 Neural network

Neural networks are available for simulating the human brain performance in remembering the information and learning. Human brain consists of a great number of nerves. Each of these nerves is connected to the other ones and sends signals to each other. Although each neuron has no the complex structure, but the set of these neurons create a more complex network. In fact, the artificial neural networks are going to create a special output by the special input and according to this, the concept of training or adjustment and learning of artificial network is achieved.

3.5 Markov model

Viewing Web transactions in the form of a series of page views allows that a number of useful models can be used to detect and analyze the user circulation patterns. One of the events is modeling the behavior of the user's circulation by Markov chain in the website. A Markov model with a set of states $\{S_1, S_2, ..., S_n\}$ and a transfer matrix is shown [6]:

$\{P_{1,1},\ldots,P_{1,n},\ldots,P_{2,1},\ldots,P_{n,1},\ldots,P_{n,n}\}$
Where $P_{i,j}$ is the probability of transition from state S_i to state S_j .

4. THE PROPOSED METHOD

In the proposed system, firstly data recorded in server weblog which are as the input data of system are preprocessed. Then, the web usage mining operations is performed on the identified sessions for building the functional patterns of user. Finally, web pages related to users' traversal method in the site are proposed based on their functional patterns. The web page is proposed by the neural Network and markov model. Following of this section will studied each of above components.

4.1 Log file preprocessing

There are various data with different formats in high volumes on the level of web. These data include HTML pages, CSS & JS files and variety of multimedia images [7]. In web usage mining, the preprocessing includes identifying users and their sessions that are used as main elements to detect the pattern [5]. An accurate identification of users and their sessions has a particular importance in web personalization because the users' models are made based on their behavior which they also themselves will be available as users' sessions.

4.2 constructing the users' session vectors

Now that we have preprocessed recorded data in the web server logs and have also achieved the proper data in the form of users' sessions, we are ready to accomplish the web usage mining.
Suppose that P equals to the set of pages accessed by users of a site according to the following definition:
$P=\{P_1,P_2,\ldots,P_m\}$
And each of the pages P_i $(1\le i\le m)$ have a specific and unique URL and S which represents the users sessions is expressed as following:
$S= \{S_1, S_2,\ldots,S_m\}$, $S_i \subset P$ $(1\le i\le m)$
Every Si session is demonstrated as an m-dimension vector:
$S_i=\{w(p_1, s_i), w(p_2, s_i), \ldots, w(p_m,s_i)\}$
In the above relation, the value of W (P_j, S_i) equals to the weight assigned to j-th visit in Si session and the value of j is among 1 to m. It should be noted that each of the above pages can be present in each of the sessions. In the above relation, the values of w must indicate the rate of users' interest in pages. One of the cases which have a relation with the user's interest in pages is the page frequency. The frequency of a page means the amount of access to that page in a session by users and it has a direct relation with the users' interest in that page. The following relationship shows the calculation of the rate of a page frequency in a session. In the following relationship, N-visited represents the number of a page visits in a session and visited pages equals to the whole set of the visited pages in a session. It is given by (1):

$$frequency(page)=\frac{N\_visited(page)}{\sum_{page\in visitedpages} N\_visited(page)} \quad (1)$$

One of the parameters that can specify the rate of user interest in a page, is the time duration spent by a user for visiting a page. The amount of time duration is a normalized value (between 1 and 0) which is calculated for each page by relation (2):

$$duration(page)=\frac{TotalDuration(page)/Length(page)}{Max_{page\in visitedpage}(TotalDuration(page)/Lengh(page))} \quad (2)$$

Another criteria which is effective on the rate of user interest in a session is the date of page visiting, this means that the pages which have been recently visited show a better reflection of user interests. But since the date of visiting is not a numerical value rather is a historical figure, some operations must be applied on it. The normalized numerical value of the date of page visiting is calculated by the following relation. In this relation, Date Origin, Date(Page) and PageLast are equal to the origin date, date of page visiting and last visited page in session, respectively. The amount of Date value is a normalized value (between 1 and 0) which is calculated from relation (3) for each page:

$$DateValue(page)=\frac{(Date(page)-DateOrigin)/(Date(pageLast)-DateOrigin)}{Max_{page\in visitedpages}((Date(page)-DateOrigin)/(Date(pageLast)-DateOrigin))} \quad (3)$$

Now, we use the harmonic mean to combine these three parameter. The value of Interest (page) that equals to the amount of user interest is obtained as (4). Let Da(page) be DateValue(page), F(page) be ferequency(page) and Dur(page) be Duration(page). So we have in (4) relation:

$$\text{int}erest(page)=\frac{3*F(page)*Dur(page)*Da(page)}{F(page)*Dur(page)+Dur(page)*Da(page)+F(page)*Da(page)} \quad (4)$$

The above value is a normalized value (between 1 and 0) that is calculated for each page.

4.3 creating the users profile

Here the user profile is an exhibitor of the resultant of his favorite pages. The users' profiles are created in this section. For doing this, first the set of sessions of all users

is separated by the separation of the user. Assume that Si ($1 \le i \le k$) is the set of the sessions of user U_i. The mean vector Sui is specified as the resultant vector for user U_i and will be indicative of the mean of user's interests in pages.
The weight of each web page in the mean vector is calculated by the average weight of that page in all the sessions of the user (S_1, S_2, ...,S_k). The user behavior history is also considered in calculating the mean vector of the session.

4.4 Clustering User Profiles and Neural Network the session

After obtaining the users profile which indicates the abstract of the users' interests in web pages, we can divide them into some groups using the clustering algorithm so that users' interests in web pages can be better organized and also it can be provided a background for extracting their internal patterns. For clustering the users' profile, we use k-mean algorithm. The result of the clustering is as k= {k_1, k_2,...,k_c} and c equals to the number of identified clusters by k-Means algorithm. Supposing that P is the set of users' profile, then we have the following formula per k_i Є k (k_i C P).

Since the weight of vectors is between 1 and 0, then their average value is also between 1 and 0. To reduce the number of dimensions, it is defined a threshold for pages weight in mean vectors of clusters. The pages whose weight is less than that threshold, are removed from the mean vector. And the remaining pages represent the most interests of users in the relevant cluster.
If we consider the set of users' movement patterns existing in website as a NP Set, then this set will be displayed as the following.
NP= {np_1, np_2,… , np_k}
In the above set, there is a mean vector for each cluster and we have for each member of the above set that each member of np_i is a subset of the set of website pages. These patterns are applied for determining the similarities between new profiles and the previous profiles.
Here, the neural network is used to find the closest cluster to the user session and to propose some proper pages to him. Network training is performed using the movement patterns obtained from the previous stages. Then, we should prepare the session of current user so that is appropriate for the neural network. Since the input of neural network is related to some weights of pages, we should create a profile for the current user based on pages weight. Now, we must determine that the new profile belongs to which one of the existing clusters. For this purpose, it is enough to give the new profile to the neural network until it can determine an appropriate number of clusters for this new profile.

4.5 Recommend pages using the Markov model in cluster

After using neural networks, was diagnosed nearest cluster to the user's current session, the next step is proposed from inside the cluster to new users, we are extracted sequences of pages that visited by Markov model. And through it propose the page with the highest Means page the most probable repeat to user. for do this we apply the markov model to training clustering data and suggest do to testing data to measure system performance, and use of the system to suggest to new users.

5. RESULT SIMULATION

In this thesis, it has been used the web server pages of Saskatchewan University for conducting the research. The data of this web server have been used as a set of web server logs which are derived from the two-week log data of the web server of the university site in 2004. This file includes 1480 user sessions and 570 pages which have been accessed by different website users during the mentioned sessions. A time period about 1800s was considered as the threshold time to extract the users sessions from the web server logs. Some pages which have been referenced less than 10 percent or more than 80 percent of the maximum frequency of access to pages were excluded according to [5]. During two stages, first 156 and then 46 website pages remained respectively, and thus 46 pages remained finally. Then, all the sessions with a length of less than 8 were removed, ultimately it remained about 617 sessions. After removing the inappropriate pages and sessions, we divided the obtained sessions into two parts. We used the first and second sessions as training sessions and test sessions, respectively. The first part is used for learning and the second part is applied for the system evaluation.
Clustering the profiles has been performed using the k-Means algorithm in the visual studio 2010 Software. Markov model has been performed in the visual studio 2010 software, and neural network has been performed in the matlab software. Thus, first the system was trained using the set of training data , then we used the test data set and created the new sessions for users by some test data sets without any role in creating the training profiles. Finally, we evaluated the efficiency of our system by these new data.

System is evaluated by two metrics including accuracy and coverage [8]. The value of accuracy is defined as a ratio of correct propositions to the whole propositions. In the other word, if the proposed equals to the set of proposed pages, R equals to the set of correct pages and size is a function

which implies the size of a set, then the value of accuracy is calculated by the (5) relation:

$$precision(proposed) = \frac{Size(precision \cap R)}{Size(R)} \quad (5)$$

Coverage is used for reviewing this issue that how many correct pages in the provided propositions have been covered by system. Its mathematical definition is as (6):

$$Coverage(proposed) = \frac{Size(proposed \cap R)}{Size(T)} \quad (6)$$

Given that the window size is usually three to four pages in papers and also we deleted the sessions with less length, then we considered the window size equal to 4 in this research.

In order to assess the proposed system based on the clustering by k-Means algorithm, we compared it with a method based on association rules and other method [9], their results have been shown in the two following figures.

Fig. 1 Compare the proposed method with other methods .

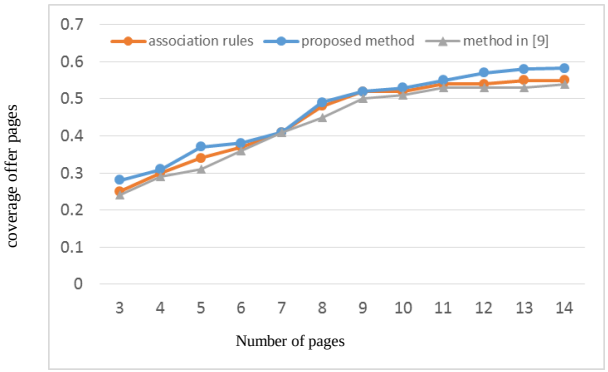


Fig. 2 Compare the proposed method with other methods.

It is completely clear that the value of accuracy has an inverse proportion to the number of proposed pages, so that the value of accuracy will reduce by increasing the number of proposed pages. This means that the number of proposed correct pages will be less than the whole pages proposed by the system.

As it is clear in the above figure, the level of coverage of both systems (proposed and based on K-Means) has a direct proportion to increasing the number of pages and also the level of coverage of both systems is increasing followed by raising the number of the proposed pages.

6. Conclusions

In this paper, a method was offered for prediction of web users' subsequent page selection using neural network and markov model. Our offered system benefits from web usage mining for recommendation to users. To achieve the users' survey pattern, a method has been used that firstly users' profile is made based on data extracted from web server records and during profile making process, history of users' behavior and page visiting date is taken into account. Then, upon clustering the profile, users' movement patterns are extracted. After obtaining movement patterns of recommendation engine using neural network and markov model, a list of suitable pages is recommended to the user. Summary of implementation shows that recommendation engine offered in this paper has appropriate accuracy and coverage for prediction of subsequent requests of user.

References

[1] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works, Journal of Expert Systems with Applications", Vol. 41, No. 4, Part 1, March 2014, pp.1432-1462.

[2] K., Santhisree, A., Damodaram, "Clustering on Web usage data using Approximations and Set Similarities, International Journal of Computer Applications", Vol. 1, No. 4, , 2010, pp.0975 – 8887.

[3] Ida Mele, "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content", ACM, WSDM'13, February 2013, pp.765-769.

[4] J., Jose, P., Sojan Lal, "Extracting Extended Web Logs to Identify the Origin of Visits and Search Keywords", Intelligent Informatics Advances in Intelligent Systems and Computing, Vol. 182, 2013, pp.435-441.

[5] G., Castellano, A. M., Fanelli, M. A., Torsello, "NEWER: A system for Neural-fuzzy Web Recommendation", journal of Applied Soft Computing, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, vol. 11, No. 1, 2010, pp.793-806.

[6] X., Dongshan, S., Junyi, "A New Markov Model for Web Access Prediction", journal of Computing in Science and Engineering, vol. 4, no. 6, 2002, pp. 34-39.
[7] K., Goseva-Popstojanova, G., Anastasovski, A., Dimitrijevikj, R., Pantev, B., Miller, "Characterization and classification of malicious Web traffic, Journal of Computers & Security", Vol. 42 , May 2014, pp.92-115.
[8] Y. S., Cho, S. C., Moon, S., Jeong, I., Oh, , K., Ho Ryu , "Clustering Method Using Item Preference Based on RFM for Recommendation System in U-Commerce", Ubiquitous Information Technologies and Applications, Vol. 214, 2013, pp.353-362.
[9] Z., khademali, A., harounabadi, J., mirabedini, "A new intelligent algorithm to creat a profile for user based on web interaction", Journal of management science letters, vol. 3, no.4, 2013, pp. 1155-1160.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Personalization Web Pages for Site Users, Utilizing Users' Interests and Sequential Patterns Discovery

Zeynab Fazelipour[1,2], Ali Harounabadi[3]

[1]MSc Student, Department of Computer, Khuzestan Science and Research Branch, Islamic Azad University Ahvaz, Iran

[2]MSc Student, Department of Computer, Ahvaz Branch, Islamic Azad University Ahvaz, Iran
Z.Fazelipour@gmail.com

[3]Department of Computer, Tehran Center Branch, Islamic Azad University Tehran, Iran
A.Harounabadi@gmail.com

Abstract

With the rapid growth of information on the Web and increase of users who are daily visiting the web sites, presenting information proportionate to requirements of users who are visiting a special website so that they could find their desired information would be essential. Therefore, analyzing browsing behavior of web users and modeling this behavior has particular importance. The aim of recommender systems is guiding users to find their favorite resources and meet their needs, using the information obtained from the previous users' interactions. In this paper, to predict the web pages with high precision, a hybrid algorithm of clustering technique, All-*K th*-Order Markov model, and neural network are presented. For this purpose, in order to model users' movement behavior, after clustering those with the same interests, the sequential patterns are extracted on users' sessions of each cluster using all-4th-order Markov model. Next, in the step of pages recommendation to a current user, which is performed in an online state, first, a current user session is assigned to a cluster using neural network. Then Markov model created on the cluster which has the nearest match to the current session, is applied and a sequence of pages, which the users are interested to view, is included in the list of recommendation. The implementation results demonstrate that the proposed algorithm has higher precision and recall comparing to other recommender systems.

Keywords: *Personalization Web Pages, Clustering, Neural Network, Markov Models*

1. Introduction

Web is an important resource for information retrieval where Increasing growth of information has caused finding the needed information to be more difficult. The main challenge users are facing is to effectively find relevant information with the minimum effort and time invested. To resolve these problems, personalization the web pages in order to customize the web environment has become a popular phenomenon. Web personalization is a process where information or services provided by a website are adapted to the needs of a user or a specific group of users by using received knowledge of user navigation behavior and his specific interests in the form of a combination with content and structure of the website can provide active suggestion according to users behavior pattern [1]. User behavior modeling is a fundamental factor in each personalized system which is done implicitly by user information or extracted users' samples [2]. Using web mining techniques in order to extract knowledge from available web information is considered one of important approaches in web personalizing and it is classified to three active researches field based on extracted web data area: web content mining, web usage mining, web structure mining [3]. So far, many personalized systems based on web-mining are created. In fact, personalization systems the web pages are an important component in a website to provide the desired or required information of the users without their explicit request. The aim of web personalization system is to recommend a series of items to the current user. Such recommendations include links, advertisements, texts, products, etc., in lines with the user's interests and preferences.

In this paper, a recommender system is presented using Markov model and based on users' interest, predicting the future requests of them. In the proposed method, after preprocessing log file and extracting of users' features from their sessions, the users profile is established in an offline state. Then, profile of users by k-means method is clustered and users' movement patterns are extracted. In

the next step, All-4 the-order Markov model is applied in order to model users' movement behaviors for each cluster. Finally, in an online state, the neural network assigns a new user to the cluster belonging to him/her and the Markov model corresponded to that cluster predicts future users' requests and gives a list of suitable pages to them. The proposed algorithm has been simulated on real-world data, and the results indicate that proposed algorithm has significantly enhanced the quality of the recommendations. The rest of this paper is organized as follows: In the second section the needed background material is provided. Section 3 includes several works done in this field Section 4 elaborates on the proposed method. The details of implementation, data set and evaluations are explicated in section 5 and finally, section 6 concludes the paper.

2. Background Material

In this section, web usage mining methods and All-K th-Order Markov model used in this paper are discussed.

2.1 Web usage mining

Web usage mining refers to process of discovering meaningful and suitable patterns of web user's application data. Application data refers to data which is stored in Web server log file by users when they use web. Web usage mining is the most applicable method for extracting users' behavior pattern in these files [4,5]. This approach concentrates on the techniques which can predict user behavior while cooperating in web. The principle duty in web usage mining is retrieve useful information from the web server records. This item is divided in to three phases. Data Preprocessing, Pattern Discovery, and Pattern Analysis [6,7].

Data Preprocessing: the collected web data usually consists of large and heterogonous information. The data should be changed to homogeneous and adaptable information which is suitable for pattern discovery phase. Like many data mining applications, preprocessing and preparing data include filling the missing values, deleting noise, changed and formatted data and eliminating incompatibility [8]. In web usage mining, this phase includes data purging, recognizing user and their session, which are fundamental factors for discovery pattern.
Pattern Discovery: In this step rules, patterns and statistic information, applying data mining techniques on user access sessions are discovered.
Pattern Analysis: seeks rules and statistic information acquired form pattern discovery phase which is very interesting for management site personnel.

2.2 All-Kth-order Markov model

Markov model is a model for studying stochastic processes which is efficient for modeling and predicting browsing behavior of website users. These models are a method for the discovery of sequential patterns in order to predict link. This modeling is based on the transition probability between web pages that are stored in the user session [9].

If $P = \{p_1, p_2, \ldots, p_m\}$ be a set of pages in a web site and W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob($p_i|W$) is the probability that the user visits pages pi next. This probability, $prob(p_i|W)$, is estimated using sequences of page visited by all users in history data, denoted by W. This is done by creating a transition probability matrix. In principle, with this method, the probability of all the pages to be the next page is calculated and then the page with the highest probability is selected as a predication.
Of course, the Markov model starts calculating the highest probability of the last page visited because during a web session, the user can only link the page he is currently visiting to the next one. But, since a precise calculation of the whole conditional probabilities is not possible, the Markov process is used for predication of the next page. This process takes a limited number of previously viewed K pages. In other words, the probability of visiting a page pi does not depend on all the pages in the web session and on a small set of k previous pages, where $k << l$ [9].

So, the next page, p_{l+1}, that the user will visit, by equation (1) is expressed [9]:

$$P_{l+1} = arg\ max\ _{p \epsilon P} \{P(P_{l+1} = p | p_l, p_{l-1}, \ldots, p_{l-(k-1)})\} \quad (1)$$

Where, k denotes the number of the preceding pages and it determines the order of the Markov model .

The Markov model order is corresponded to the number of prior events in predicating a future event. Hence, the Markov model from k-th order predicts the probability of next event by viewing the previous K event. For calculate probabilities, if S_j^k be a state containing k pages, $S_j^k = \{p_{l-(k-1)}, p_{l-(k-2)}, \ldots, p_l\}$. The probability of $P(p_i|S_j^k)$ from a history (training) data set, using equation (2) is estimated [10]:

$$P(p_i|S_j^k) = \frac{Frequency(\langle S_j^k, p_i \rangle)}{Frequency(S_j^k)} \quad (2)$$

In most cases, the Markov models with lower order (first or second) cannot exactly predict the next page that is to be visited by user. This is because these models do not review profoundly the user's history. Thus, in order to achieve to a

better precision, the Markov models with higher order (third or fourth) must be used. But, higher-order Markov models have a number of limitations associated with high state-space complexity, low coverage, and sometimes even worse prediction accuracy. One method to overcome the problem of low coverage is to train varying order Markov models and then combine them for prediction. This scheme is called the All-K th-order Markov model.

Utilization of all-K th order Markov model usually contributes to produce separate models for each K-order. If the model is unable to predict by k-th order, it will try to do it by reducing the model order gradually. In this scheme, for each test instance, the highest-order Markov model that covers the instance is used for prediction. For example, if the first, second, and third-order Markov models is built, then, given a test instance, first a prediction is done using the third order model. If this model does not contain the corresponding state, then a prediction is done using the second-order model, and so on [11].

3. Review of the Previous Works

In [12] have proposed an experimental system which classifies users of browsing patterns using a combination of web usage and content mining. In this paper, firstly, the profile of users is created based on information achieved from web server logs. Subsequently, browsing pattern of each class of users is derived applying clustering algorithm to profiles. Afterwards, achieved result is combined with content of corresponding web page to generate users' browsing pattern for predicting future requests. At the final stage, recommender engine generates a list of user desired pages using neural network.

In [13] has proposed a hybrid model to improve web page prediction. The main goal of this paper is optimizing efficiency of Markov model by K-NN classification algorithm. This hybrid system consists of two parts. In the first one, a training dataset is selected and is classified into multiple classes that classes illustrate data pattern on data set. In the second part, Markov model is applied on classified data in order to predict next page.

In [14] has provided a mixed system for discovery and analysis of users' movement patterns. In this paper, after preprocessing log file data, a clustering algorithm based on ants for pattern discovery is used which is out of system line phase. In the online phase, movement patterns are classified using decision tree classification method and user's next request in a web site is predicted.

In [15] has suggested a method for predicting web access pattern. This model is a combination of Boosting and Bagging methods that improves the accuracy of the prediction model. It aims to create a prediction model for behavior of user's random browsing so that the next requested pages could be estimated according to previous visited pages.

Construction steps prediction model provided with Bagging, Boosting as follows:

- Data Preprocessing (Data cleaning, Session identification)
- Apply Bagging method on the training data set
- Apply Boosting method on the training Data set
- Combine bagging and boosting prediction results

The results show that the accuracy of the Bagging and Boosting prediction is improved as compared with Markov Model and Markov Model combined with ARM.

In [16], using users' behavior analysis, an approach, which makes use of preprocessing of log-file, is presented in order to pre-fetch, predict, and improve the performance of the web server. Clustering, Markov Model, and association rules are the three techniques used in recommending web pages. Therefore, after preprocessing and identifying the sessions, they are clustered using K-means clustering algorithm and measuring similarities. Every data-set is grouped in a different cluster. Then, Marco Model predicts the results. In case of ambiguity, association rules are applied to present accurate results. The knowledge base in this system is a reservoir of features which are mined using data-mining techniques. These features include the number of users, the visited web pages, and the time to access the pages.

4. Proposed Method

The proposed method in this paper is based on web usage mining approach. The proposed algorithm consists of two phases: offline phase and online phase. In the offline phase, at first web server registries are preprocessed, and then the users' sessions are extracted. Then, Users profile is created using a mean vector based on each user's interest and with K-means clustering method; the resulting profiles are clustered so that the users' movement patterns to be extracted. In the proposed system, the All-4th order Markov model is used to model users' movement behaviors in each cluster. Then, in online phase, the movement pattern (cluster) corresponding to an active user session are indentified using neural network and finally using predictive process of the corresponded Markov

model, the sequence of pages are predicted for users. The proposed algorithm includes the following steps:

1. Preprocessing of the server registries in order to extract the user's sessions. Preprocessing of the server registries include data cleaning, user identification, and session identification.

2. Dividing the obtained sessions into two categories of training and test sessions set.

3. Extraction of users' characteristics and weighting the viewed pages during sessions

4. Making users' profiles using training sessions

5. Apply k-means clustering method in order to cluster profiles and extract movement users' patterns

6. Modeling sessions of each cluster using all- 4th-order Markov model

7. Training the neural network using obtained movement patterns

8. Active user simulation using test sessions set

9. Utilization neural network in order to determine the relevant cluster using test sessions set. If the current session is placed in the relevant cluster, the algorithm continues, otherwise it goes back to step 5.

10. Recommend a list of predicted pages to the current user using Prediction Process by Markov model corresponding to detected cluster.

4.1 Determining Weight of Pages

When the users' sessions are indentified, the redundant pages are deleted. The reason is that if the number of pages is beyond the normal limit, the clustering process needs too much time. Therefore, in the proposed method, we use the effective dimension reduction method in order to improve the clustering outcomes. For this purpose, inspired by [17], we omit the pages whose degree of support (a ratio of the number of sessions containing that page to the total pages) is too low or too high. The pages with poor support are those that are rarely visited and not viewed over total sessions. These pages have little informational value and not suitable to include in the movement patterns. In contrast, the pages with strong support are those that are almost always viewed during sessions. They are homepages of a site that are displayed in most of sessions. Thus, the pages are deleted when access to them is below 10% over the maximum accesses. Those pages viewed above 80% are also omitted, like the home pages. Then, all the user sessions with length less than 5 pages are deleted because they cannot be predicted by 4-th order Markov model.

We represent each session s_i as an m dimensional vector over the space of web pages, $s_i = \{w(p_1, s_i), w(p_2, s_i), \ldots, w(p_m, s_i)\}$, where $w(p_j, s_i)$ is a weight assigned to the jth web page $(1 \leq j \leq m)$ visited in the session si. In this paper, the weight $w(p_i, s_j)$ is defined as the interest degree of a particular user to the page, which is the harmonic mean of page observation time and page observation frequency to represent this interest [12].

4.2 Creating Users' Profile

This system module is used to create users' profiles. For this purpose, session vectors related to different users separated. Assuming $\{s_1, s_2, \ldots, s_k\}$ be a set of session vectors of ith user (u_i). We compute a mean vector sui for the user u_i as its representation. This mean vector represents web pages, which are interesting in by the users. The weight of each web page in the mean vector is computed by the average weight of the web pages across total access sessions of the user $\{s_1, s_2, \ldots, s_k\}$.

4.3 Clustering Profiles

In order to cluster the obtained profiles, the K-means clustering algorithm is used. The following algorithm is a basic one for this method.

1- First, K points are selected as center points of clusters.

2- Each data case is assigned to the cluster whose center has the shortest distance to that data.

3- Calculation of new centers of each cluster. Each new center will have its own mean value of clustered data points.

4- The above steps 2 and 3 are reiterated until no variation occurs in the cluster centers any more.

The relation (3) is expressed as an objective function [18]:

$$SSQ(N, C) = \sum_{i=1}^{k} \sum_{X \in N_i} d(X, C_i) \qquad (3)$$

Where, $d(X,C_i)$ is a measure of distance between the points and C_i is ith cluster center.

After apply the k-means clustering method to the obtained profiles, several cluster centers are resulted as $C = \{c_1, c_2, \ldots, c_m\}$ in which every c_i $(1 \leq i \leq k)$ indicates a subset of the users' sessions, and k indicates the number of clusters. We

compute a mean vector (m_c) for each cluster $c \in C$ as its representation. The mean value for each web page in the mean vector is equal to the mean of the weight of that page to total sessions that cluster. Every average vector shows the browsing pattern of users in a cluster in a special class of accessed web pages. As the results of profiles clustering, *NP = {np1, np2, ...,npk}* is used to represent the set of users navigation patterns, in which each *npi* is a subset of *P*, the set of web pages.

4.4 Creation Markov Model on Sessions of Each Cluster

After clustering the profiles, all-4th order Markov model is applied to model the users' movement sessions for each cluster. In this way, on each cluster, the transition probability matrix is established between web pages for 1-4th order Markov model. To create a transition probability matrix in the Markov model, the value of transition probability is estimated between web pages in the sessions using the relation (2) that is introduced in the section (2-2).

4.5 Finding Corresponding Cluster Utilizing the Neural Network

In online mode, we use neural network to find the most similar cluster to the user's current access session. Therefore, we train neural network using users' navigation patterns. The navigation patterns have been considered as the inputs of the network and the relevant cluster's number as the output of the one.
Neural network input is a vector of weights of web pages visited in the session. So, a profile for the users' current session based on the weight of the pages is created. Now, should be determined that current session profile belongs to which cluster (navigation pattern). For this purpose, current session profile is made, to the neural network input are given and the network determines relevant cluster's number for the session.

4.6 Prediction of Web Pages Using All-4th-order Markov Model

In this step, having identified the current session cluster, the all-4th-order Markov model related to that cluster predicts the preferred page for users. For each test session, the transition probability matrix of 4th order Markov model related to that cluster is employed to predicate the current page corresponding to the test session. The pages with the highest transition probability are included in a list of recommended pages. If each state is not covered by 4th order matrix, the 3, 2 and 1th order matrices are matched so that the sequence of pages that the user is interested to view is included in the recommendation list.

5. Implementation and Evaluation of the Proposed Method

To implement the components of the proposed system, Microsoft SQL Server and MATLAB software were utilized. The log file data of this research is collected based on Nasa log file web server. After preprocessing and identifying sessions, in order to train and test the proposed system, the 70% of sessions were selected for training and the rest were used for test and evaluation. Then, the proposed system is taught using the educational data. The optimal number of cluster compactness and cluster separation proposed in [19]. The clustering method applied and produced 7 clusters. A perceptron network was used to learn from the data. After training the system, the experimental data are used which were not responsible in making the movement patterns, and simulated the active user. The aim of personalization is to calculate a proposed set (rs) for the user's current session which has the highest correspondence with the user's interests. This part is the only online component of the system and must be highly efficient and precise.

5.1 Evaluation Metrics

Two criteria of precision and recall are effective parameters in system performance that using formulas (4) and (5) are obtained. Precision refers to the capability of recommender system to generate precise recommendations. In other words, the precision of recommendation equals to the ratio of correct recommendations to total number of recommendations. Recall refers to the ability of the recommender system to generate suggestions which could be seen by the user. In fact, recall is the ratio of diagnosed correct recommendations to remained pages in the continuation same session [20].

$$Precision(rs, rp) = \frac{|rs \cap rp|}{|rs|} \quad (4)$$

$$Recall(rs, rp) = \frac{|rs \cap rp|}{|rp|} \quad (5)$$

Where, rs is the output of the proposed system (suggested set) and rp is pages viewed by the user in the continuation same session.

5.2 Comparing the Proposed Algorithm with other Methods

In this section, In order to evaluate suggested method as in Figures 1, 2 and 3 are shown, first, we compared the precision of our method with association rule Markov model [21], the integrated Markov model (second order) with association rule and clustering [10] and the hybrid Markov model (third order) with clustering [22] for predication of next web page. The precision of next page prediction is equal to percentage of the number of correctly predicted sessions to total test sessions.

As well as, the precision and recall of the proposed algorithm are compared respectively, with the performance of the recommender systems NEWER [17] and IPACT [23] than the number of suggested different pages with the window length of 4.
The experimental results demonstrate that the proposed algorithm has higher precision and recall than algorithms that were compared.
As it can be seen in Figure (1), the proposed method outperformed the other methods.

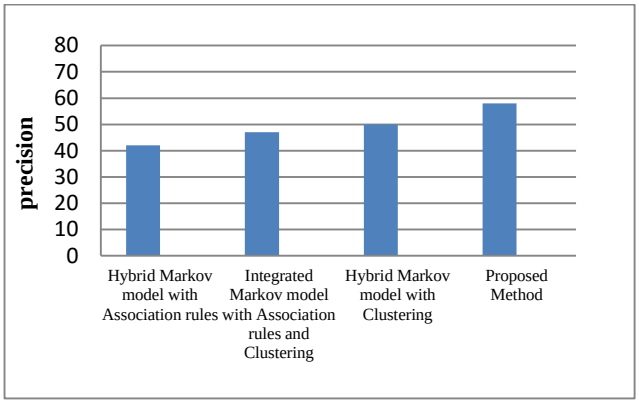


Fig. 1 Comparison of the precision algorithms for predicting the next web page

As it can be seen in Figure (2), except in the number of proposed page of 3, in which NEWER method has higher accuracy than the proposed algorithm, in other parts the proposed algorithm is more accurate compared to the other two methods.

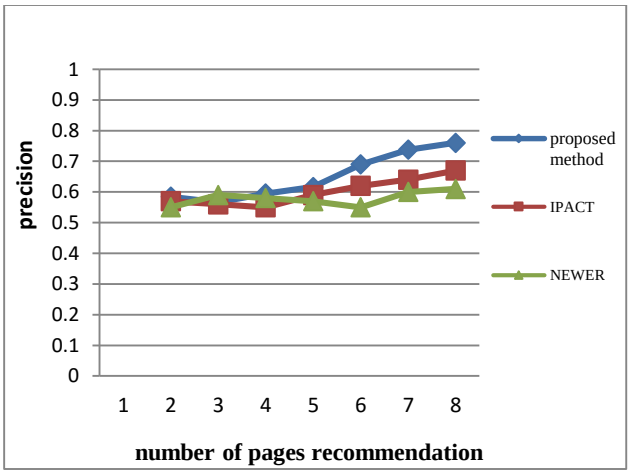


Fig. 2 Comparison of the precision algorithms for the number of suggested different pages

Figure (3) indicates the recall of the algorithms being compared with one another. By increasing the number of proposed pages, the proposed algorithm, as the figure (3) shows, has a greater recall in comparison with the NEWER method except about size of suggested pages 3, 5 and 6. Such improvement is also true compared with the IPACT method except in the range of 2 to 4 pages.

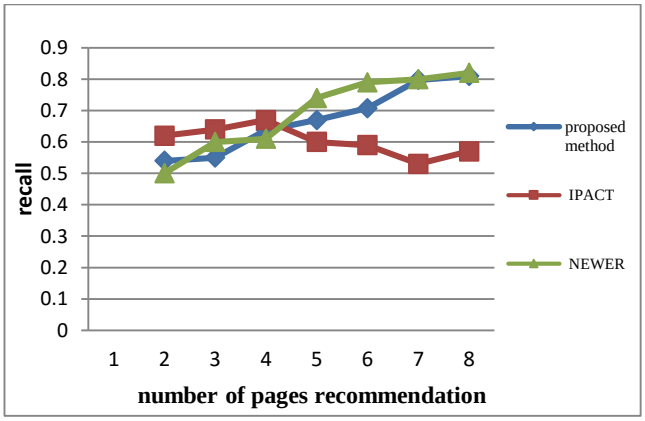


Fig. 3 Comparison of the recall algorithms for the number of suggested different pages

6. Conclusions

In this paper, a new hybrid algorithm of clustering technique, All-*Kth*-Order Markov model and neural network was presented. In this algorithm, the profiles of users who have similar movement behaviors are clustered

using K-means clustering method. Then, to model users' movement behavior, sequential patterns were extracted from the users' sessions of each cluster using all-4th-order Markov model. Next, in the step of pages recommendation that was done in the state of online, first, a current user session is assigned to a cluster using neural network. Then, the Markov model established on that cluster, which has the nearest match to the current session, is applied, and the sequence of pages which the user desires to view is included in the recommendation list. It can be concluded that a better efficiency of the system results from modeling and prediction of higher order Markov Model with all orders. Also, the analysis of the model was done on users' cluster which could model the users who had the same interests and produce more precise prediction using the more homogenous and limited sessions. Evaluation results revealed that the proposed method has high precision and recall while recommending pages to users.

References

[1] Eirinaki, M., Vazirgiannis, M. ," Web mining for web personalization," ACM Transactions on Internet Technologies (TOIT), NY, USA, Vol. 3, No. 1, 2003, pp. 1-27.

[2] Forsati, R., Meybodi, M., "An Algorithm Based on Structure of Connected Pages and Information of Users for Suggesting Web Pages", The Second Iran Data Mining Conference, industrial Amir Kabir university, 2008.

[3] Chen, P.Z. and Sun, C.H. and Yang, S.Y., " Modeling and Analysis the Web Structure Using Stochastic Timed Petri Nets", Journal of Software, Vol. 3, No. 8, 2008, pp. 19-26.

[4] Tyagi, N.K. and Solanki, A.K. and Wadhwa, M., "Analysis of server log by web usage mining for website improvement," International Journal of Computer Science Issues (IJCSI), Vol. 7, No. 8, 2010, pp. 17-21.

[5] Verma, V., Verma, A.K. and Bhatia, S.S., "Comprehensive analysis of web log files for mining," International Journal of Computer Science Issues (IJCSI), Vol. 8, No. 6, 2011, pp. 199-202.

[6] Chitraa, V. and Davamani, A.S., " A survey on preprocessing methods for web usage data," International Journal of Computer Science and Information Security, Vol.7, No. 3, 2010, pp.78-83.

[7] Santra, A.K. and Jayasudha, S., "Classification of web log data to identify interested users using naïve Bayesian classification," International Journal of Computer Science Issues (IJCSI), Vol. 9, No. 2, 2012, pp. 381-387.

[8] Pamutha, T., Chimphlee, S., Kimpan, C., Sanguansat, P., "Data preprocessing on web server log files for mining users access patterns," International Journal of Research and Reviews in Wireless Communications (IJRRWC), Vol. 2, No. 2, 2012, pp. 92-98.

[9] Thwe, p., "proposed approach for web page access prediction using popularity and similarity based pagerank algorithm", international journal of scientific & technology research (ijstr). Vol. 2, No. 3, 2013, pp. 240-246.

[10] Khalil, F., H., Li, Wang, H., "An Integrated Model for Next Page Access Prediction", International Journal knowledge and Web Intelligence (IJKWI), Vol.1, No.1/2, 2009, pp.1-18.

[11] Deshpande, M. and Karypis, G., "Selective Markov models for predicting Web Page accesses," ACM Transactions on Internet Technology (TOIT), Vol. 4, No. 2, 2004, pp 163-184.

[12] Rashidi, S.F., Harounabadi, A., Abasi Dezfouli, M., "Prediction of Users Future Requests Using Neural Network", Management Science Letters. Vol. 2, No. 6, 2012, pp. 2119-2124.

[13] Kaushal, P. ,"Hybrid markov model for better prediction of web page," International Journal of Scientific and Research Publications(IJSRP), Vol. 2, No. 8, 2012, pp. 1-4.

[14] Sujatha, V., Punithavalli , "An approach to user navigation pattern based on ant based clustering and classification using decision tress," International Journal of Advanced Engineering Sciences And Technologies, Vol. 1, No. 2, 2010, pp. 112 – 117.

[15] Girija P., Kavitha V., "An Approach for Predicting User's Web Access Pattern", International Journal of Computer Science and Management Research, Vol.2, No.5, 2013, pp. 2585-2589.

[16] Makker, S., Rathy. R.K, "Web Server Performance ptimization using prediction pre fetching Engine", International Journal of Computer Applications, Vol. 23, No. 9, 2011, pp. 19-24.

[17] Castellano, G., Fanelli, A.M., Torsello, M.A., "NEWER: A System for Neuro-fuzzy Web Recommendation", Applied Soft Computing, Vol. 11, No.1, 2011, pp. 793-806.

[18] Jain, A. K., Murty, M. N. and Flynn, P. J., "Data Clustering: A Review," in: ACM Computer, Vol. 31, No. 3, 2000, pp. 264-323.

[19] Liu, H., Keselj, V., "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," Data and Knowledge Engineering, Elsevier, Vol. 61, No.2, 2007, pp. 304-330.

[20] Nakagawa, M., Mobasher, B. ,"A hybrid web personalization model based on site connectivity," In Web KDD Workshop at the ACM SIGKKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 59-70.

[21] Khalil, F. and Li, H. and Wang, H.,"A framework of combining Markov model with association rules for predicting web page accesses," Proc. of the 5th Australasian Conf. on Data Mining (AusDM'06), Australian Computer Society, 2006, Vol. 61, pp. 177-184.

[22] Pandey, T., Kumari, R., Tripathy, A., and Sahu, B., "Merging data mining techniques for web page access prediction: Integrating Markov model with clustering," International Journal of Computer Science (IJCSI), Vol. 9, No.1, 2012, pp. 188-193.

[23] AlMurtadha, Y. and Sulaiman, M., Bin, N., Mustapha, N. and Udzir, N.I., "IPACT: Improved web page recommendation system using profile aggregation based on clustering of transactions," American Journal of Applied Sciences, Vol. 8, No. 3, 2011, pp. 277-283.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A Method for Optimizing Maintenance and Querying Ontology-based Linked Data

Naghmeh Sohrabian[1], Bita Shadgar[2]

[1] Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University, Ahvaz, Iran
sohrabian@isc.gov.ir

[2] Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University, Ahvaz, Iran
bita.shadgar@scu.ac.ir

Abstract

At present, emerged technologies such as Resource Description Framework (RDF) are used to describe information in the semantic web. RDF triples are the basic components of linked data, which build the whole structure of the semantic web. Alongside the semantic web development, RDF data are also growing in scope and volume rapidly. As a result, the size of T-Boxes and also A-Boxes in linked data-related ontologies is undergoing a great change. The scale of ontology-based linked data requires efficient structures for storing and also querying on these data. This paper proposes a method based on relational databases for storing ontology-based linked data. This method achieves shorter query response time and more accuracy comparing other known RDF storage methods such as schema-oblivious, schema-aware and hybrid methods. To evaluate the results, DBpedia infobox ontology and dataset has been used.

Keywords: Linked Data, Ontology, Relational Database, Resource Description Framework, Indexing.

1. Introduction

Linked data come from different domains in various data sources on the semantic web. RDF links interlink these data and therefore in a near future make the whole semantic web connected. RDF triples are the basic components of linked data. They consist of three parts: subject, predicate and object. Subjects and predicates are identified with a unique global identifier named URI and object values can be URIs or literals. In recent years, linked data have grown so much in scope and volume [1]. DBpedia data source which converts Wikipedia data to the suitable format for the semantic web, is the nucleus for the semantic web [2] and it alone consists of billions of linked data available in about 100 languages. These data need efficient structures for maintenance and retrieval in a way that query response time and storage space size be acceptable.

This paper introduces a method for mapping ontology basic components to relational database components. It uses relational database for both linked data and ontology storage. It is desirable to store large ontologies with related instances in relational databases. Because Relational databases have long been used as primary sources for semantic web data and also ontology storage. They have also ensured the best facilities for storing, updating and querying the data from different domains [3-6] and they reduce the barriers for data exchange and integration. Furthermore using Relational Databases permits web application to query via SQL (Structured Query Language) instead of SPARQL (Simple Protocol and RDF Query Language), the semantic web query language which is not as matured as SQL in supporting the operations needed for querying data. SQL is relationally complete [5-7]; this means that any relational algebra operation such as select, projection, join and union can be modeled with SQL. SQL provides query capabilities using Data Manipulation Language (DML) and schema definition capabilities using Data Definition Language (DDL) [8].

Mapping ontologies to relational databases consists of three steps: schema mapping, data mapping and query mapping. Schema mapping builds the relational database schema based on the source ontology T-Box; data mapping converts RDF data to the relational tuples and query mapping translates SPARQL queries to SQL [9,10].

The rest of this paper is structured as follows. Section 2 shortly introduces existing methods for storing RDF data in relational databases along with their strong and weak points. Section 3 describes the proposed method. Section 4 compares the proposed method query response time with the other methods for different test queries and finally, conclusions and future works are discussed.

2. Related methods

Currently, there are several methods for mapping between RDF data model and relational databases [10,11]; however, all of them have some drawbacks, or are intended for certain purposes. These methods fall into four groups: (1) schema-oblivious method, (2) schema-aware method, (3) data-driven method and (4) hybrid method.

2.1 Schema-oblivious (also called generic or vertical)

One ternary relation (table) is used to store RDF triples. This table contains triples of the form <subject-predicate-object>. Fig. 1 shows the structure of the table. Different properties of a specific resource are tied together using the same subject URI. Attribute “subject” represents a resource that is the source of property, the property name is given in attribute “predicate” and attribute “object” represents a destination resource or literal value for the property. Well-known Schema-oblivious RDF stores include Jena [12,13],

Sesame[14], KAON [15], RStar [16] and OpenLink Virtuoso [17].

triples

subject (resource URI)	predicate (property name)	object (property value)

Figure 1: Schema-oblivious storage method

2.2 Schema-aware (also called specific or binary)

This approach usually employs ontology to generate equivalent property relations and class relations in relational databases. Unlike the previous representation, one table per RDF/S schema property or class is used. A property table, Property(s,o), is created corresponding to each property in ontology and then stores each subject s and object o which are related by this property. A class relation, Class(i), is created for each class in ontology and stores instances i of this class. Fig. 2 shows the schema of the relational database. Representatives of schema-aware RDF stores are Jena [12,13], DLDB [18], RDFSuite [11], DBOWL [19], and PARKA [20]. This method considers a datatype proportionate to the type of datatype property in the related ontology.

$Property_1$

subject (resource URI)	object (property value)

$class_1$

subject (resource URI)

…

$Property_n$

subject (resource URI)	object (property value)

$class_m$

subject (resource URI)

Figure 2: Schema-aware method

2.3 Data-driven

This method uses RDF data instead of RDF schema or ontology, to generate database schema. For instance, database schema can be generated based on the patterns found in RDF data using data mining techniques. Property relations are created when their instances are first seen in an RDF document during data mapping. This method is seldom implemented in storage systems. It is used by sesame [14].

2.4 Hybrid

This method uses the combination of the features of the schema-oblivious and schema-aware methods. In this method, a schema-oblivious database representation is partitioned into multiple relations based on the data type of object o. So, property/class instances with range values of the same type are stored in the same relation and a binary relation, Class(i, c), is introduced to store instances i of classes c. Fig. 3 displays the relational database schema for this method.

Properties with range resource

subject (resource URI)	predicate (property name)	object (property value)

Class instances

subject (resource URI)	object (classid)

Properties with range integer

subject (resource URI)	predicate (property name)	object (property value)

Figure 3: Hybrid storage method

3. Method Description

The proposed method builds a storage system in which most kinds of related queries are answered in a relatively short time and reasonable storage space with more accuracy comparing the previous methods. Linked data and the ontology related to them are the inputs of the proposed system and relational database schema with ontology instances that are stored in relations are the output. This system uses DBpedia dataset infobox data and ontology[1] in order to test the proposed model. Fig. 4 shows the general structure of the proposed system. As Fig. 4 shows, the method consists of three main steps. The first and the primary one is transformation of ontology T-Box to relational database schema. In the second step, relational database schema is constructed based on DDL commands which have been generated in previous step. In the third step, relational tables are filled with linked data extracted from the dataset. These data are available in N-triples format.

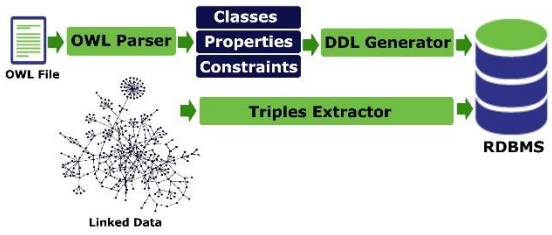


Figure 4: General structure of the proposed system

[1] Available for download at http://wiki.dbpedia.org/data-set-37

3.1 Translation of ontology T-Box to relational database schema

After validating the syntax of this file, ontology file is decomposed to its structural components such as classes, object properties, datatype properties and constraints. For each of these components separate DDL commands are generated in order to build the equivalent structure in relational database schema.

3.1.1 Transformation of ontology classes to relational structures

Fig. 5 shows the general steps for extraction of classes from the source ontology and generation of DDL commands for building proper relational structures. breath-first search is applied to the ontology file initially. Owl ontology class definitions are recognized with <owl:class> elements. First of all, owl:thing class, which is the parent of all ontology classes is added to a queue. After that, in each hierarchy level, classes are observed one after another and their attributes take proper values. For any class definition in ontology file, the class attributes such as rdf:about, rdfs:label, rdfs:subclassof and rdfs:comment are given proper values. DDL command corresponding to create a relational database is the first command to be written in the output file. After that, another DDL command is written for generating the meta table named classes with the schema seen in Fig. 6. This meta table stores the general information for any class in ontology. Attribute state stores “non-leaf” in case of root classes and “leaf” in case of “leaf” classes. When class definition search in ontology ends, DDL commands to fill table classes with proper data are written.

3.1.2 transformation of object/datatype properties to relational structures

This step is somewhat similar to the previous one. Here object and datetype properties definitions in ontology file are used to fill meta table named Property. They are recognized with <owl:objectproperty> and <owl:datatypeproperty> elements respectively. The first DDL command is written to generate meta table properties in relational database schema. This table stores the meta data for ontology properties. Fig. 7 shows the schema for this table. Domain and range attributes store URIs of the source and target of each property. As OWL does not contain any data type itself, it uses data types from XML schema. Attribute flag is used to distinguish between object and datatype properties. In the proposed method, MySQL 5.5 is used as RDBMS. Therefore, data types in OWL ontology file should be mapped to proper data types in MySQL. Similar to hybrid method, property tables are categorized based on object’s data type. Fig. 8 displays their schema.

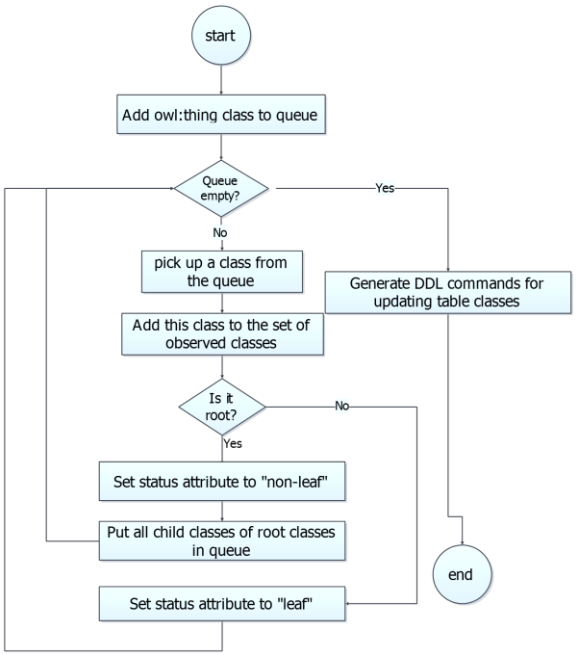


Figure 5: Transformation of ontology classes to relational structures

classID	title	URI	state	comment

Figure 6: the schema of table classes

propertyID	title	URI	domain	range	flag

Figure 7: The schema of table properties

statementID	subject	predicate	object

Figure 8: The schema of property tables for storing instances

These relations are the most important ones. They are used to store triples whose predicates are datatype properties and types_resources table is used to store triples whose predicates are object properties. For each triple the property(predicate) range specifies where to store the triple.

3.1.3 Application of ontology relations to database schema

In this step, all the relations in the source ontology are transformed to relational structures. Based on the input ontology there exists various kinds of relations. Storing these relations can be useful while inferencing new data from existing RDF triples. In DBpedia ontology, relations come in four groups: rdfs:subclassof, owl:equivalentclass, owl:equivalentproperty, owl:functionalproperties. For subclasses, a table with two columns is generated: one for class URI and another for parent class URI. In order to

store equivalent classes and equivalent properties two tables with two columns are generated which store class/property URI and equivalent class/property URI. Functional properties are stored in a single column table. One of the strength points for the proposed method is the generation of a meta table named propertyClass_instances. As Fig. 9 shows, for each property this table stores a unique identifier named ID, property URI, classID which points to a class in table classes and flag which distinguishes between object and datatype properties. Another attribute named table_ is the name of the table which is going to contain RDF statements of a specific property in the next step. Therefore, there is no need to include all property tables in queries. Instead, only the retrieved tables are used for querying in proposed approach. Using this table facilitates and accelerates the queries which ask for the instances of a particular property. Querying the individuals of a particular class is the same story. Again, the attribute table_ value is used as a reference for the storage table containing RDF triples. So, this table also facilitates the queries on all RDF statements that are related to a specific class.

ID	property	classID	flag	table_

Figure 9: The schema of propertyclass_table

The proposed method generates another table named resourceClass which links each resource to the class which it belongs to. This table has two columns: one stores resource URIs and another one stores class URIs. This table facilitates and accelerates the queries which ask for the parent classes of each resource.

3.2 Generation of relational database

Before loading linked data into relational tables, the whole schema of relational database should be built. Executing DDL commands which are generated in previous steps builds the ontology T-Box. Moreover, meta tables classes and properties are filled with proper RDF data.

3.3 Filling relational database with extracted linked data

The entry to this step is the relational database schema that has been generated in previous step. Here the RDF triples are converted to the relational tuples depending on the target table. In DBpedia dataset, linked data are stored in infobox properties and infobox specific properties parts. Objects are stored as simple or typed literals. Therefore, they include extra texts such as language labels, XML data type URIs and some extra characters such as “”, “<”, “>” and so on. To extract related RDF triples out of this dataset, objects of the datatype properties should be modified in a way that these extra texts are removed and the genuine object is retrieved. To find the proper table for storing each RDF triple, the value for attribute table_ is used.

4. Results Evaluation

In this section, after application of all these methods to DBpedia dataset, the query response time and storage space are compared with the proposed method (with or without indexing) in case of queries with different viewpoints: queries on linked data structural components such as subject, predicate, object, queries, queries on resources’ parent classes and queries on class-related linked data. Then, the storage space of the proposed method is compared with the other methods.

4.1 Capability of response to different query types

4.1.1 Queries which ask for linked data subjects

In schema-oblivious method, the below SQL command retrieves RDF statements having a particular subject identified with the subject URI:

```
select subject,object,predicate from triples where subject=@subjectURI
```

In schema-aware method, all property tables should be searched which is very slow and inefficient. Because for each property table, a union operation is added to the SQL query. The SQL query generated is as follows:

```
select subject,object,predicate from property1 where subject=@subjectURI
union
select subject,object,predicate from property2 where subject=@subjectURI
union
.
.
union
select subject,object,predicate from propertyn where subject=@subjectURI
```

$property_1$ to $property_n$ are the first and last property tables which contain RDF triples respectively. In hybrid method, always the same number of tables is explored. This number depends on the number of data types which are defined in ontology. In case of DBpedia infobox ontology this number equals 11. The SQL command is as follows:

```
select subject,object,predicate from types_1 where subject=@subjectURI
```

union
.
.
union

select subject,object,predicate **from** types_11 **where** subject=@subjectURI

The proposed approach first retrieves the property table names that contain the specified subject via an SQL query as follows:

(1) **select** **distinct** table_ **from** propertyClass_instances, propertyClass_table **where** propertyClass_instances.classID= propertyClass_table.classID **and** resource=@resourceURI

Then, a union operation is applied to the retrieved tables:

(2) **select** subject,object,predicate **from** types_1 **where** subject=@subjectURI
union
.
.
union
select subject,object,predicate **from** types_n **where** subject=@subjectURI

As the number of queried tables decreases, the query speed increases in comparison with the previous method. In order to increase this speed even more, an index is added to subject column in all property tables. Query speeds in each case are evaluated with three random URIs in DBpedia dataset. Fig. 10 shows the results for this kind of query in terms of time. The parameter for the query1 is the first URI, for the query2 is the second URI and for query3 the third one. As seen in Fig. 10, the proposed method response time is decreased about 47 percent in case of query1, 44 percent in case of query2 and 40 percent in case of query3. When indexing is applied to column subject, response time decreases 35 percent comparing the situation without indexing in case of query 1, 59 percent in case of query 2 and 60 percent in case of query 3.

4.1.2 Queries which ask for linked data predicates

In schema-oblivious method, the below SQL command retrieves linked data predicates.

select subject,object,predicate **from** triples **where** predicate=@propertyURI

In schema-aware method, a simple SQL query retrieves the specified data:

Select subject,predicate,object **from** property[x]

property[x] is the property table which contain the specified data. This method is very efficient in response to queries of this type.

Hybrid method behaves the same as querying on subjects, but instead it asks for predicates.

The proposed approach retrieves the names of property tables which contain the specified predicate:

select **distinct** table_ **from** propertyClass_table **where** property=@propertyURI

Then, a union operation is applied to retrieved tables:

select subject,object,predicate **from** types_1 **where** predicate=@propertyURI

union
.
.
union

select subject,object,predicate **from** types_n **where** predicate=@propertyURI

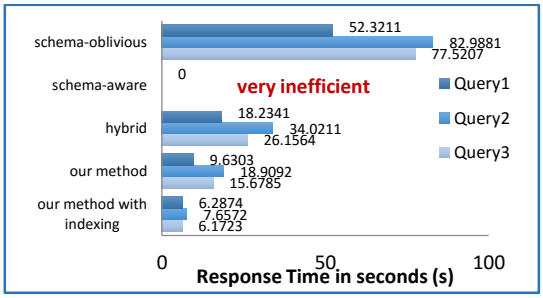


Figure 10: Comparison of response time for query on linked data subjects in seconds (s)

So, similar to the previous condition with decrease in number of queried tables, query speed increases. Additionally, applying index to predicate columns increases this speed even more. It should be notified that the proposed method infers new triples from the main ones and adds them to the retrieved results. For this purpose, the similar properties to the queried property are searched using table sameProperties. However, to avoid increasing the storage space this method does not store inferred triples in any structure. Instead, it adds them to the main triples in run time. Fig. 11 shows the results of query responses in this case.

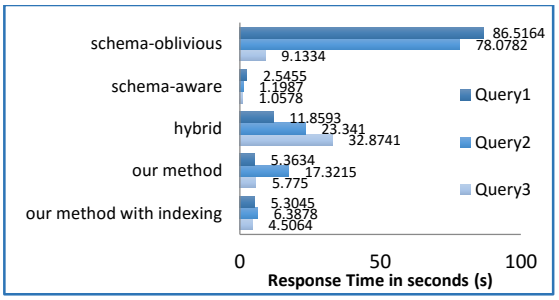


Figure 11: Comparison of response time for query on linked data predicates in seconds (s)

4.1.3 Queries which ask for linked data objects

This kind of query needs two parameters: property URI and specific range for property values. So, here just the data property case is studied. In schema-oblivious method, no SQL query can extract the object value out of the third part of RDF triple. In schema-aware method, a SQL query similar to the one on predicates in this method retrieves the objects in special range of values. Again, the method applies union operation to so many tables and therefore results in bad query results. Hybrid method behaves the same as querying on subjects and predicates, but instead it asks for predicates. The proposed approach here is similar to the previous one, but here the objects are queried in specific ranges. Fig. 12 shows the time results of all methods in case of queries on objects. In order to evaluate the time performance of queries, three random URIs with random ranges are selected. The results show that when there is no indexing, schema-aware method performs the best. But totally, the proposed method with indexing is the best in terms of time. It shows 43 percent decrease in time in case of query1, 8 percent in case of query2 and 82 percent in case of query3.

4.1.4 Queries on resource parent classes

As schema-oblivious method lacks the class information, this kind of query cannot be applied to this method. Both schema-aware and hybrid contain structures for storing the instances of a specific class, but lack the possibility of querying on the parent classes that are related to a resource. The proposed approach uses table resourceClass to ask for class instances.

select classURI **from** resourceClass **where** resource=@classURI

For each individual, the proposed method only stores leaf classes in ontology tree. After retrieving a particular class URI it refers to table subclasses to append all the parent classes for the specified resource to the list of retrieved classes. Furthermore, it queries table sameClasses to find the equivalent classes with the ones that are retrieved as resource parent classes and adds the results to the previous retrieved classes. Finally, the proposed method adds an index on column resource to increase the query speed. Fig. 13 shows the results of this query in terms of time for three random class URIs.

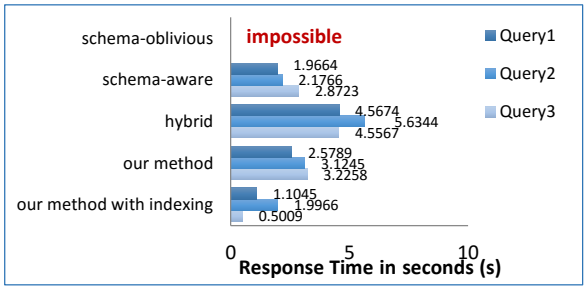


Figure 12: Comparison of response time for query on linked data objects in seconds (s)

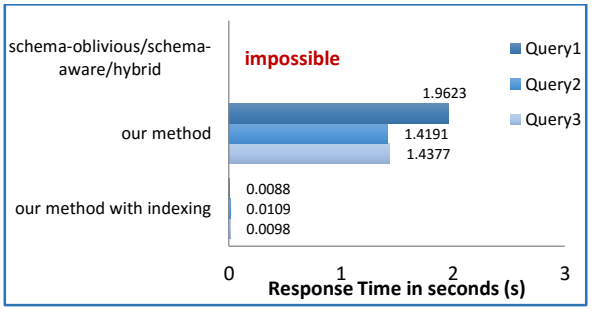


Figure 13: Comparison of response time for query on resource parent classes in seconds (s)

The results show that indexing decreases query response time about 99 percent in case of query1, 80 percent in case of query2 and 99 percent in case of query3.

4.1.5 Queries on class-related linked data

This kind of query retrieves all linked data that are related to a particular class in the form of RDF triples. As schema-oblivious method lacks the class information, this type of query is not possible to execute in this method. In schema-aware and hybrid method the SQL query which retrieves class-related linked data is as follows:

select subject,object,predicate **from** class[x],property$_1$ **where** class[x].resource=property$_1$.subject
union
.
.
union
select subject,object,predicate **from** class[x],property$_n$ **where** class[x].resource=property$_n$ .subject

n is the number of properties. As there are so many join and union operations, this method time performance is very inefficient. The proposed method uses table

propertyClass_instances. First a SQL query on table propertyClass_instances retrieves all the property and target table names for storing the RDF statements that are related to a specified class:

```
select property,table_ from propertyClass_instances where class=@classURI
select subject,predicate,object from types_1 where predicate=@propertyURI
union
.
.
union
select subject,predicate,object from types_n where predicate=@propertyURI
```

So, the proposed method retrieves all the class-related linked data without joining any tables. This increases the query speed. Then, an index is added to column predicate which increases query speed even more. Fig. 14 shows the response time for this type of query in all methods.

4.2 The storage space

The storage space of different methods can be investigated here from two points of view: number of generated tables and the volume of relational database.

4.2.1 Number of generated tables

As the number of generated tables increases, the storage and retrieval overhead also increases. Furthermore, data management and updates get harder. The schema-oblivious method uses just one table for the storage of RDF statements. All extracted linked data are stored in this table. As previously mentioned, this method lacks any structure for the storage of ontology properties and classes. Application of this method to the DBpedia infobox dataset results in about 14,000,000 RDF triples being stored in one table. The schema-aware method considers a table for each class or property in ontology. Application of this method to dataset results in generation of 314 tables for classes, 851 tables for object properties and 893 tables for datatype properties. Hybrid method generates a table for each class to store the instances of that class. For each group of property data types, a table is generated to store the RDF triples. Application of this method to dataset results in generation of 314 tables for the storage of classes and 10 tables for the storage of datatype properties. The proposed method does not consider any structure for the storage of classes. It generates one table for the storage of object properties, 10 tables for datatype properties and 8 meta tables for storing general information. Table 1 represents the number of tables for each method.

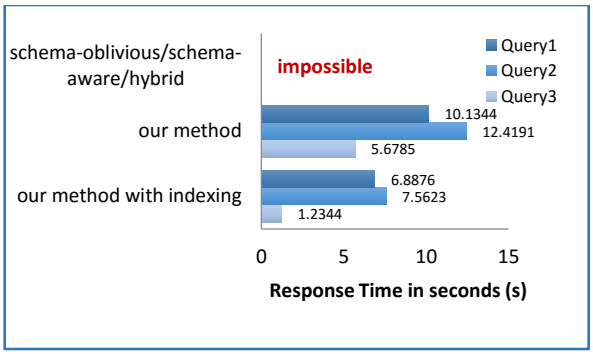


Figure 14: Comparison of response time for query on class-related linked data in seconds (s)

Table 1: The number of generated tables

Storage method	Number of relations
schema-oblivious	1
schema-aware	2058
hybrid	325
our method	**18**

4.2.2 The relational database volume

Obviously, as the volume of database increases, the storage and retrieval overhead also increase. Table 2 shows the total database volume for each method. It shows that the proposed method generated database is the lowest in volume.

Table 2: The total generated database volume in Gigabyte

Storage method	Database Volume
schema-oblivious	2.2 Gigabyte
schema-aware	3.92 Gigabyte
hybrid	2.45 Gigabyte
our method	**2.13 Gigabyte**
our method with indexing	2.23 Gigabyte

5. Conclusion

In previous section the response performance of queries on ontology-based linked data and the storage volume for existing methods such as schema-oblivious, schema-aware and hybrid methods and proposed method are investigated and compared. The results show that the schema-oblivious method is only efficient in response to queries which ask for subjects, predicates or objects of Linked data. This method lacks any structure for the storage of classes or properties. It stores all RDF triples in one table. This causes problems with query speeds when performing some operations like joining the table with itself. Schema-aware generates tables for any class or property in ontology. This

causes an overhead on whole the database and makes data management and updates hard. Furthermore, for most of the queries so many union operations should be included. This method performs well just in response to the queries which ask for class individuals or instances of a particular property. Hybrid method performs well in response to queries which ask for specific range of values in addition to the query types which are supported by schema-aware method. Hybrid method has resolved the problem with the number of generated tables in schema-aware method, but it still contains all property tables in some queries. The proposed method aims at resolving the problems with the previously discussed methods. Furthermore, it can respond well to all the query types which are mentioned in this article. It uses indexing on queried data column to speed up the queries and uses inference to increase the accuracy of the retrieved RDF data. Furthermore, the number of generated tables is independent of the number of classes in ontology. The results show that in most of the cases, the proposed method with indexing performs the best in terms of response time, result completeness and simplicity of queries which are used to retrieve data and it supports most types of queries comparing the other methods.

References

[1] C. Bizer, F. Universitat, T. Heath, T. Berners-Lee, "Linked data - the story so far", *International Journal on Semantic Web and Information Systems*, Vol. 5, No. 3, pp. 1-22, 2009.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, "DBpedia: a nucleus for a web of open data", in *The 6th International Semantic Web Conference (ISWC2007)*, Busan, Korea, November 2007.

[3] M. d. M. Roldan-Garcia, J. F. Aldana-Montes, "A Survey on Disk Oriented Querying and Reasoning on the Semantic Web", *in The 22th IEEE ICDE Workshop SWDB*, Atlanta, 2006.

[4] D. Beckett, J. Grant, "SWAD-Europe Deliverable 10.2: Mapping Semantic Web Data with RDBMSes", (last modified: 23 January 2003), [accessed: 11 April 2012], <http://www.w3.org/2001/sw/Europe/reports/scalable_rdbms_mapping_report/>.

[5] D. E. Spanos, P. Stavrou, N. Mitrou, "Bringing Relational Databases into the Semantic Web: A Survey", *Semantic Web*, Vol. 0, No. 0, pp. 1-41, 2010.

[6] I. Astrova, N. Korda, A. Kalja, "Storing OWL Ontologies in SQL Relational Databases", *in proceedings of the World Academy of Science, Engineering and Technology*, 2007.

[7] D. E. Spanos, P. Stavrou, N. Mitrou, "Bringing Relational Databases into the Semantic Web: A Survey", *Semantic Web*, Vol. 0, No. 0, pp. 1-41, 2010.

[8] C. J. Date, *An Introduction to Database Systems*, 8th edition, Addison Wesley , Boston, 2003.

[9] W3C Incubator Group, "A survey of current approaches for mapping of Relational Databases to RDF", (last modified: 8 January 2009), [accessed: 5 April 2012], <http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf >.

[10] A. Chebotko, S. Lu, X. Fei, F. Fotouhi, "RDFPROV: A relational RDF Store for querying and managing scientific workflow provenance", *Data & knowledge Engineering*, Vol. 69, No. 1, pp. 836-865, 2010.

[11] Y. Theoharis, V. Christophides, G. Karvounarakis, "Benchmarking Database Representation of RDF/S Stores", *in Proceedings of the 4th International Semantic Web Conference(ISWC2005), LNCS 3729*, Galway, Ireland, 2005.

[12] K. Wilkinson, C. Sayers, H. Kuno, D. Reynolds, "Efficient RDF Storage and retrieval in Jena2", *in the first International Workshop on Semantic Web and Databases*, Berlin, Germany, 2003.

[13] B. McBride, "Jena: Implementing the RDF Model and Syntax Specification", *in proceedings of the second international workshop on semantic web(semweb2001)*, Hong Kong, China, 2001.

[14] J. Broekstra, A. Kampman, F. V. Harmelen, "Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema", *in Proceedings of the first International Semantic Web Conference(ISWC2002)*, Chia, Sardinia, Italy, June 2002.

[15] T. Gabel, Y. Sure, J. Voelker, "KAON – An Overview: Karlsruhe Ontology Management Infrastructure", *University of Karlsruhe*, 2004.

[16] L. Ma, Z. Su, Y. Pan, L. Zhang, T. Liu, RStar: An RDF Storage and Query System for Enterprise Resource Management, *In proceedings of the International Conference on Information and Knowledge Management (CIKM)*, Washington, DC, USA, 2004.

[17] O. Erling, Implementing a SPARQL compliant RDF triple store using a SQL-ORDBMS, Technical report, OpenLink Software Virtuoso, 2001, Available from http://virtuoso.openlinksw.com/wiki/main/Main/VOSRDFWP.

[18] Z. Pan, J. Heflin, DLDB: Extending Relational Databases to Support Semantic Web Queries, *In Proceedings of the International Workshop on Practical and Scalable Semantic Web Systems (PSSS)*, Sanibel Island, Florida, USA, 2003.

[19] S. Narayanan, T. M. Kurc, and J. H. Saltz. DBOWL: towards extensional queries on a billion statements using relational databases. Technical Report OSUBMI_TR_2006_n03, Ohio State University, 2006. Available from http://bmi.osu.edu/resources/techreports/osubmi.tr.2006.n3.pdf.

[20] K. Stoffel, M.G. Taylor, J.A. Hendler, Efficient Management of Very Large Ontologies, *In proceedings of the American Association for Artificial Intelligence Conference(AAAI)*, Palo Alto, California, 1997.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Social Impact on Android Applications using Decision Tree

Waseem Iqbal [1], Mohammad Irfan[2] and Muhammad Asif[3]

[1] Department of IT & CS , University of Sargodha, Gujranwala Campus
Gujranwala, Punjab, Pakistan
waseem@canwascomputers.com

[2] Department of IT & CS , University of Sargodha, Gujranwala Campus
Gujranwala, Punjab, Pakistan
arfan.uosgrw@gmail.com

[3] Department of IT & CS , University of Sargodha, Gujranwala Campus
Gujranwala, Punjab, Pakistan
masif.uosgrq@gmail.com

Abstract

Mobile phones have evolved very rapidly from black and white to smart phones. Google has launched Android operating system (OS), based on Linux targeting the smart phones. After this, people became addicted to these smart phones due to the facilities provided by these phones. But the security leaks possess in Android are the big hurdle to use it in a secured way. The Android operating system is mostly used because it is an Open Source/freeware and most of its applications are also freely available on different online applications stores. To install any application, we must accept the terms and conditions regarding the access to multiple part of device and personal information, otherwise unable to install these free or paid applications. The main problem is that when we allow the access to multiple parts of our device and our personal information, the inherited security leaks become more vulnerable to threat. A very simple and handy solution is that we only install the applications that are positively reviewed by other users who already installed and are still using these applications. We implement the Decision Tree, a machine learning technique, to analyze these positively reviewed application and make a recommendation whether to install them in the device or not.

Keywords: *Android, Decision tree, Machine Learning Technique, Social Impact, Entropy*.

1. Introduction

Mobile devices have become essential part of our life. These mobile devices, especially mobile phones have evolved very rapidly [1] from a simple mobile phone with a black and white display to smart phones with color display. These smartphones not only make usual phone calls and SMS's but also read documents, create presentation, enjoy audios and videos, play games, and surfing the internet [2]. The Google introduced "Android" as its first open platform operating system for mobile devices, in2007 [3]. Now android has become the largest operating system for mobile devices and on every passing day, more than one million devices worldwide are being activated on android [4]. Android is extensively used open source operating system for mobile devices under the Open Handset Alliance [5] which makes it compatible with several hardware (devices and architectures) and software (applications). The popularity of Android has also been increased due to availability of source code with no cost [6]. Now android has become a standard for hardware manufacturer as well as software developers. Android provides you a world-class platform for creating apps and games for android users everywhere, as well as an open marketplace for distributing to them instantly [7].

The Android operating system (OS) was built on Linux Kernel [6] and specially designed to run on mobile devices. In spite of gaining popularity with every passing day [8], these mobile devices are also facing numerous security threats [9] [10].

In August 2010, a report from Essential Security against Evolving Threats (ESET) security systems shows the past five months description in which 65% of the threats were reported and if we categorize these threats, 30% are available for download in different markets, 37% spread through SMS and 60% threats are transferred through one device to another [11].

In year 2012, two departments of USA, the department of justice and the homeland security also issued a report in which these departments have shown that 79 percent of mobile OS malware threats over the whole world took

place over the android platform while the 0.7 percent threats have been detected over iPhone Operating System (IOS) platform [12]. The public sector organizations suggested their workers not to use the android due to its security threats. The report indicated that android is the main target for such attacks due its enormous market shares, free of cost behavior and open-source architecture [12].

In 2014, another report from Kaspersky labs stated that 14900 new malicious programs have been added in database of Kaspersky in 1st quarter of 2012 while this figure increased by three times in 2nd quarter of the same year [13]. The reason behind this huge number of attacks is that the android devices are based on Linux which is an open source and can easily be exploited to create different malicious applications. These applications can easily bypass the android permissions system to complete the installation process.

A very large number of applications (freely or for some cost) are available on different online applications stores like Google Play Store, 1 Mobile Market and many others. These applications have become the vital part of mobile devices and without these applications these devices lose their importance. The applications, for example, Facebook, Skype, WhatsApp, Viber, MS Office 365 and many others are the beauty of smartphones and due to these applications the usage of smartphones has increased. These and many other available applications are very user friendly and make the user more productive and connected with others at any-time, anywhere. Users just only view the application and try to install the application without knowing the security threats about the application. Whenever a user wants to install an application, this process requires an access to the multiple parts of device and personal information and if user denies granting the access, he or she will not be able to install the application [14].

Many other system [15] [16] [17] [18] have been devised that work on intrusion detection systems but they do not analyze the social impact on the application which is a very easy way to differentiate a trusted application from a non-trusted one. However, some research works [19] [20] have also analyzed the social impact in another way. This paper mainly describes a very handy and easy to access information for checking the social impact of any software with the help of Decision Tree. Every application, whenever anyone tries to install, gives information of total downloads, user reviews, ranking given by the users and many other information.

As we know that the bulk of data or applications are available on the internet but due to the enormous number of users, all the data or applications are reviewed or commented by the users [19] [20]. These reviews/comments are equipped with the rating in term of five stars or numbers from 0 to 5. These rating are very helpful in a way that the concerned data/application has an overall behavior among the current users.

In this paper, we have inspected the usefulness to classify the applications on the basis of available information by applying machine learning technique like decision tree [21]. This inspection is based on certain aspects. When anyone tries to download an application, the system checks available application's information and decides whether to install this application or not.

As part of the effort to prevent the threat happening, we are trying to help the users install these applications in a way by categorizing these applications either positive or negative on the basis of information available on application stores from current users of these applications. We are introducing a proactive approach to install the applications on the basis of available information which can reduce the chances of threat.

2. Proposed Solution

In this paper, we have worked on available information on application stores especially the total number of downloads, ranking in the form of stars or in the form of numbers from 1 to 5 and the recent feedbacks or reviews of the users. It is very convenient to use this information because there are huge numbers of reviews available even for a recently launched application.

The reviewers of any application often recapitulate their overall opinion about that application in the form of ratings and this information will be used for Decision Tree. So, we do not need any kind of manual data for application evaluation purposes. Same type of research work has been found in [22].

As discussed earlier, there are several online application stores but the data source, selected in this paper, is the Google Android store which is a very huge database for this purpose and contains thousands of applications with hundreds and thousands of reviews even for a very fresh application available in app store. Another reason is that the Google app store ratings or reviews are in the form of stars and numbers from 0 to 5 which can be converted or used in numerical values for

analysis purpose. The analysis result is in the form of positive or negative which recommends whether an application should be installed or not. In this paper, Decision Tree has been used for analysis purpose.

2.1 Decision Tree

Decision Tree is a best choice to apply in this scenario because other algorithms do not have ability to generate proper results accordingly because of some imperfections inherited in them. The other algorithms are List-Then-Eliminate, Find-S and Candidate Elimination methods are not suitable in this scenario due to some of these listed limitations [23]:

2.1.1 List-Then-Eliminate Algorithm:

- This algorithm can only be applied when we have finite Hypothesis Space (H).
- This algorithm requires to find all hypothesis in H, which is an improbable approach.

2.1.2 Find-S Algorithm:

- This algorithm cannot tell whether it has learned concept or not.
- This algorithm cannot give information about any inconsistency in the training data.
- This algorithm only selects the most specific hypothesis h, that's why is called Find-S.
- A complete hypothesis space may contain more than one consistent hypothesis but this algorithm only selects the most specific one.

2.1.3 Candidate Elimination Algorithm:

- The Candidate-Elimination algorithm covers many limitations of the List-Then- Eliminate and Find-S algorithms but still it has some shortcomings in it:
  - This we want to classify a new instance, it can only be classified if all the selected consistent hypotheses (Version Space) agree on the classification.
  - This target concept (say c) should exist within the hypothesis space H.
  - This algorithm does not have disjunctive learning ability.
  - This algorithm can only identify noisy data.

Decision Tree approximately resolves all these stated problems, however it has its own limitations. But in so far our scenario is concerned, it is the best one because we have discrete value attributes on which we decide whether to install the application or not. In decision tree, each inner node evaluates the attribute while each branch corresponds to the attribute value and at the end leaf nodes represent the classification for new instances. This representation has built in conjunction and disjunction properties, for example each path has conjunctions of attribute evaluation while Decision Tree itself is disjunctions of these paths (conjunctions). The Decision Tree is Preference Biased which means that the hypothesis space is complete (target function must be there) but the search is incomplete (missing some attributes while constructing the Decision tree). The basic Decision Tree has been implemented by the algorithm ID3 which has a drawback that we cannot back track it. The same algorithm has been adopted in many other research studies like [24]. The proposed model used for our research analysis is:

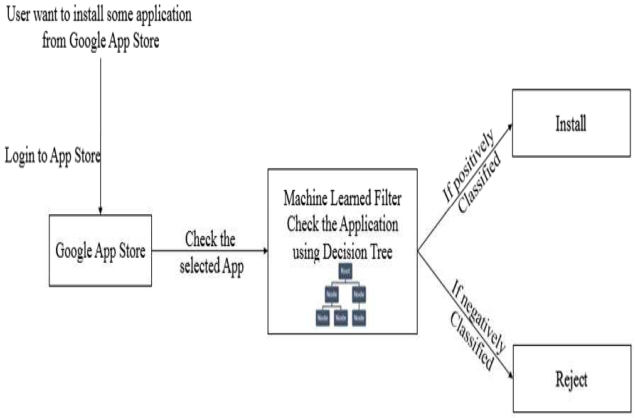


Fig. 1. The Abstract Model

Following 15 training examples have been taken to construct the decision tree:

Table 1. Real time Training Data for 15 applications with their classifications

S. No.	App Name	Google App Store	Rating	Rating in latest 30 comments	Application Downloads	Classification
1	Face Book Messenger	Yes	4.1	122	1000000000	Positive
2	Latest Android	No	4.2	104	200000	Negative
3	Whats App Messenger	Yes	4.4	110	5000000000	Positive
4	Subway Train Rush	Yes	3.7	111	1000000	Positive
5	360- Antivirus Security Free	Yes	4.5	117	5000000000	Positive
6	Free Password	No	3.5	68	250000	Negative
7	Simulator Laser Game	Yes	3	42	1000000	Negative
8	Blurb Check Out	Yes	2.1	41	100000	Negative
9	Bubble Blast	No	4.8	106	150000	Negative
10	Rabbit Rush	No	3.5	53	10000000	Negative
11	Temple Run 2	Yes	4.3	107	1000000000	Negative
12	Candy Crush Saga	Yes	4.4	111	1000000000	Positive
13	Electric Screen WallPaper	Yes	3.4	110	5000000	Negative
14	Toilet & Bath Room Rush	Yes	3.6	101	1000000	Negative
15	Mobile Hacker	No	4.8	123	20000000	Negative

2.1.4 Selection of Attributes as nodes in Decision Tree:

In this paper, the ID3 algorithm has been used for constructing the Decision Tree. The ID3 algorithm tests attributes to place the nodes at each level of Decision Tree. The finest way to find the best suitable attribute for a node or root node is Information Gain, a statistical quantitative measure, which defines how well a selected attribute splits the training examples according to the target function. The calculation of Information Gain (IG) assists in selecting the attribute to classify the examples at each level, the root node has been selected on the basis of maximum information gain from all attributes.

Table 2. Training examples distribution

Positive Examples	Negative Examples	Total Training Examples
10	05	15

Now calculate the entropy, which describes the impurity of an arbitrary collection of examples S, in our scenario the collection comprises fifteen training examples. For the collection S that contains positive and negative examples of target concept, the entropy relative to this Boolean classification is 0.9182 using the following formula:

$$\text{Entropy}(S) \equiv -p^{\delta} \log_2 p^{\delta} - p_{\varphi} \log_2 p_{\varphi}$$

Eq. (1)

The calculation for Information Gain, from collection of training examples S and attribute A, has been done using the formula:

$$\text{Gain}(S,A) \equiv \text{Entropy}(S) - \sum_{V \in \text{Values}(A)} \frac{|Sv|}{|S|} \text{Entropy}(Sv)$$

Eq. (2)

Table 3. Overall Information Gain of all attributes

Attributes	Entropy	Information Gain (IG)	Ranking based on IG
Google app store	0.9182	**0.2515**	1st
Ratings	0.9182	0.00	4th
Latest 30 comments	0.9182	0.1893	2nd or 3rd
Downloads	0.9182	0.1893	2nd or 3rd

From the above table, it is clear that the Google app store is the best classifier and selected as the root node on the basis of maximum Information Gain. After selecting the root node, now move further to select the other attributes as sub-node(s) of root nodes on the bases of IG. The entropy of Google app store is 1 based on training examples. The IG values of other attributes under the Google app store are given as:

Table 4. Information Gain of attributes based on Google app store

Attributes	Entropy	Information Gain (IG)	Ranking based on IG
Ratings	1	0.1081	2nd or 3rd
Latest 30 comments	1	**0.2365**	1st
Downloads	1	0.1081	2nd or 3rd

Now the level 2 node has been selected based on maximum IG calculated from the remaining attributes.

Table 5. Information Gain of attribute based on Latest 30 comments

Attributes	Entropy	Information Gain (IG)	Ranking based on IG
Ratings	0.994	0.1104	2nd
Downloads	0.994	**0.1832**	1st

The attribute under the Latest 30 comments attribute has also been selected based on maximum IG calculated from the remaining attributes and placed at level 3. The only remaining attribute is rating, which is placed at level 4. The following decision tree has been constructed according to the above information:

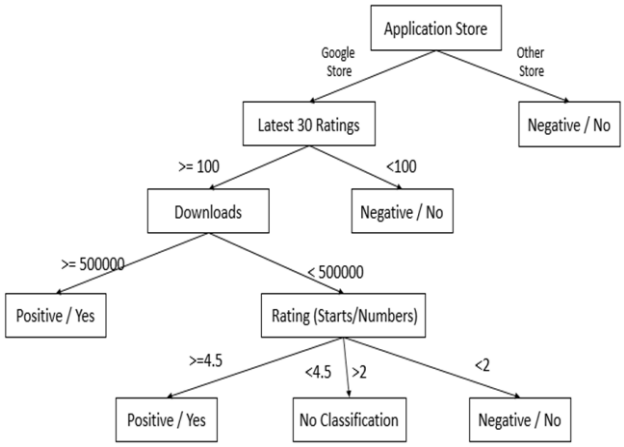


Fig. 2. Decision Tree before pruning

But this decision tree has an issue that is, the examples which have Google app store is yes, Latest 30 ratings >= 100, Downloads < 500000 and Rating is between 4.5 and 2 (terminal values are excluded) have not been classified. For decision making process, it is necessary to apply the post pruning technique of decision tree so that all examples are classified as positive or negative. As there are more negative examples for rating between 4.5 to 2, so, all the examples are negatively classified. The final decision tree after post pruning is as under:

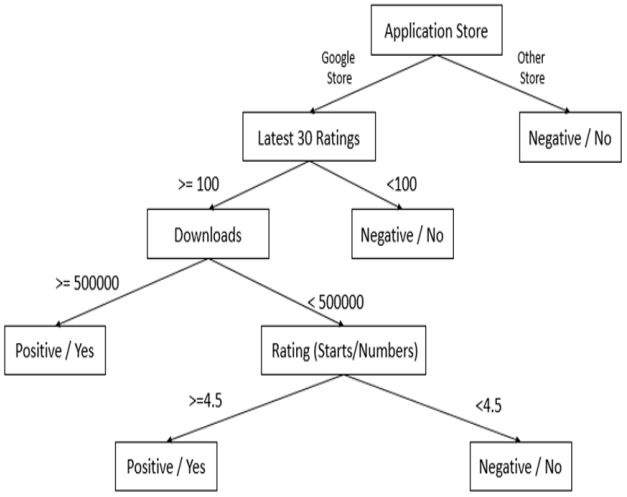


Fig. 3. Decision Tree after pruning

The Pseudo code for the final decision tree is:

Level 1. Check the availability from Google App Store

Application would be further checked at Level 2
Application would not be installed

Level 2. Check the latest 30 comments
Aggregate rating >=100 (out of 150)

Application would be further checked at Level 3

Application would not be installed

Level3. Check the numbers of downloads >=500000

Application would be further checked at Level 4
Application would not be installed

Level 4. Check the overall application rating >= 4.5
Application would be classified positively and installed
Application would not be installed if the overall rating < 4.5

After checking the given testing examples on the final constructed decision tree, following results have been found:

Table 6. Real time Testing Data for 15 applications with their classifications

No.	App Name	Google App Store	Rating in latest 30 comments	Application Downloads	Classification
1	PTCL TV	Yes	122	1000000	Positive
2	Film Actors HD Pics	No	72	10000000	Negative
3	Chrome Browser Google	Yes	96	5000000000	Negative
4	Pepi Skate 3D	Yes	102	1000000	Positive
5	Fingerprint Thermometer	Yes	107	10000000	Positive
6	Wifi Password Hacker	Yes	51	1000000	Negative
7	Speed VPN	Yes	105	1000000	Positive
8	Go Weather Forecast	Yes	104	5000000000	Positive
9	Subway Rusher 3D	No	111	10000000	Negative
10	Top Racing Car Game	Yes	108	5000000	Positive
11	Nimbuzz Messenger	Yes	96	10000000	Negative
12	Thumb Password	No	99	250000	Negative
13	Geo TV	Yes	102	10000000	Positive
14	NOAA Weather Radio	Yes	63	10000	Negative
15	Free TV	No	132	10000000	Negative

All the testing examples are classified correctly. These values can be adjusted to discrete values if there is any restriction for example the 4.5 can be rounded to 5 etc. Also application markets other than Google app store are not reliable and most likely to be vulnerable to the threats that are why we reject all other types of markets in our example.

3. Conclusion

The proactive approach, described in this paper, cannot prevent all the threats but can reduce the probability of threat happening. There are always back doors or security holes even in a very secure OS and we are always trying to mitigate them but cannot stop them all. In the same sense, we apply a machine learning technique that filters the malicious program on the basis of social impact (previous and current comments or feedback from users) and recommends whether to install the application or not.

4. Future Work

Until now, we have applied the Machine Learning Technique (Decision Tree) which decides on the basis of given stars or numbers rating by the users. In future, our aim is to apply an extra layer to filter the applications by analyzing the written comments through Natural Language Processing (NLP). The users not only give stars or numbers rating but they also give description which contains more precise knowledge about the behavior of applications.

References

[1] A. AH and R. M, "Device-aware desktop web page transformation for rendering on handhelds,"Personal and Ubiquitous Computing, vol. 9, no. 6, pp. 368-380, 2005.
[2] J. A. Chow GW, "A Framework for Anomaly Detection in OKL4-Linux Based Smartphones," in *6th Australian Information Security Management Conference*, 2008.
[3] J. DiMarzio, Android A Programmers Guide, McGraw-Hill Osborne Media, 2008.
[4] http://developer.android.com/about/index.html, http://developer.android.com. [Online]. [Accessed December 2014].
[5] W. Enck , O. Machigar and D. Patrick, "Understanding Android Security," *IEEE security & privacy*, vol. 7, no. 1, pp. 50-57, 2009.
[6] A. Shanker and L. Somya , "Android porting concepts," in *IEEE International Conference on Electronics Computer Technology (ICECT)*, 2011.
[7] "http://developer.android.com/guide/basics/whatis-android.html," Google, [Online]. Available: http://developer.android.com. [Accessed 8 June 2014].
[8] Gartner, "Gartner Says Worldwide Mobile Phone Sales Declined 8.6 Per Cent and Smartphones Grew 12.7 Per Cent in First Quarter of 2009," Gartner, Egham, UK, 2014.
[9] B. Sun, Z. Chen, R. Wang, F. Yu and V. Leung, "Towards adaptive anomaly detection in cellular mobile networks," in *IEEE Consumer Communications and Networking Conference*, 2006.
[10] B. Sun, Y. Xiao and K. Wu, "Intrusion Detection in Cellular Mobile Networks," in *Wireless Mobile Network Security*, Springer, 2007, pp. 183-210.
[11] BORTNIK and SEBASTIÁN, http://www.welivesecurity.com/2011/12/20/2012-predictions-more-mobile-malware-and-localizedattacks/, ESET, 20 December 2011. [Online]. [Accessed 14 October 2014].
[12] A. Majumdar, "http://tech.firstpost.com/newsanalysis/ us-warns-government-workers-aboutandroid- malware-threats-104558.html," 27 August 2013. [Online]. [Accessed 11 December 2014].
[13] M. La Polla, F. Martinelli and D. Sgandurra, "A Survey on Security for Mobile Devices," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, vol. 15, no. 1, pp. 446-471, 2013.
[14] A. P. Felt, E. Chin, S. Hanna, D. Song and D. Wagner, "Android permissions demystified," in *ACM conference on Computer and communications security*, 2011.

[15] D. Damopoulos, S. A. Menesidou, G. Kambourakis, M. Papadaki, N. Clarke and S. Gritzalis, "Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers," *Security and Communication Networks*, vol. 5, no. 1, pp. 3-14, 2012.
[16] Y. Zhang, W. Lee and Y. A. Huang, "Intrusion detection techniques for mobile wireless networks," *Wireless Networks*, vol. 9, no. 5, pp. 545-556, 2003.
[17] M. Miettinen, P. Halonen and K. Hatonen, "Hostbased intrusion detection for advanced mobile devices," in *IEEE conference on Advanced Information Networking and Applications*, 2006.
[18] A. Shabtai, U. Kanonov and Y. Elovici, "Intrusion detection for mobile devices using the knowledgebased, temporal abstraction method," *Journal of Systems and Software*, vol. 83, no. 8, pp. 1524- 1537, 2010.
[19] A. Girardello and F. Michahelles, "Explicit and Implicit Ratings for Mobile Applications," *GI Jahrestagung*, vol. 1, pp. 606-612, 2010.
[20] A. Girardello and F. Michahelles, "AppAware: Which mobile applications are hot?," in *ACMIinternational conference on Human computer interaction with mobile devices and services*, 2010.
[21] K. N. and H. C. C. , "Input feature selection for classification problem," *IEEE Trans on Neural Networks*, vol. 13, no. 01, pp. 143-159, 2002.
[22] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Association for Computational Linguistics*, 2002.
[23] T. M. Mitchell, Machine Learning, New York: McGraw-Hill Comp., Inc., 1997.
[24] A. Kalpesh, G. Aditya, D. Amiraj, J. Rohit and H. Vipul, "Predicting Students Performance Using ID3 and C4.5 classification Algorithms," *International journal Data mining and knowledge management process*, vol. 3, no. 5, 2013.

Waseem Iqbal is currently working as IT Manager at Canwas Computers, Gujranwala. He is MSCS scholar at University of Sargodha, Gujranwala Campus. His main interests are Machine Learning and Artificial Intelligence.

Mohammad Arfan is currently working as Network Administrator University of Sargodha, Gujranwala Campus. He is MSCS scholar at same University. His main interests are Wireless Network, Mesh Networks and Artificial Intelligence.

Muhammad Asif is currently working as a Lecturer at University of Sargodha, Gujranwala Campus. He is MSCS scholar at same University. His main interests are Machine Learning and Operating System.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Analysis of the "Heroes of the Storm"

Shuo XIONG[1], He ZHAI[2] , Long ZUO[3], Mingyang WU[4] and Hiroyuki Iida[5]

[1,2,3,4,5] School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, 923-1211, Japan
[1]xiongshuo@jaist.ac.jp
[2]zhaihe@gmail.com
[3]zuolong@jaist.ac.jp
[4]gawain513@jaist.ac.jp
[5]iida@jaist.ac.jp

Abstract

Game refinement is a unique theory that has been used as a reliable tool for measuring the attractiveness and sophistication of the games considered. The refinement measures were derived from game information progress model and have been applied in various types of games. This paper focuses on the game refinement theory and its application to a MOBA game "Heroes of the Storm" (HOS), which was produced by Blizzard Entertainment in 2014. Furthermore, we evaluate the measurement for different maps of HOS. Experimental results show that a game refinement value of HOS was between 0:08~0.1 for which previous works have confirmed.

Keywords: *Game refinement theory, Heroes of the Storm, MOBA.*

1. Introduction

1.1 MOBA game

Multi-player Online Battle Arena (MOBA) [1], also known as Action Real-Time Strategy (ARTS), in which a player controls a single character at one of two teams. The objective is to destroy the opponent team's main structure with the assistance of periodically spawned computer controlled units. Player characters typically have various abilities and advantages that improve over the course of a game and that contribute to a team's overall strategy. A fusion of action games and RTS games, players do not construct either buildings or units [4]. The genre traces its roots to Aeon of Strife (AOS), a custom map for StarCraft where four players each controlling a single powerful unit and aided by weak computer-controlled units were put against a stronger computer-controlled faction [5]. Defense of the Ancients or "DOTA" [2], a map based on Aeon of Strife for Warcraft III: Reign of Chaos and The Frozen Throne, was one of the first major titles of its genre and the first MOBA for which has been kept sponsored tournaments. It was followed by two spiritual successors: "League of Legends" (LOL) and "DOTA 2". Generally, the Original MOBA game map is shown in Figure 1. From the Figure 1, we can see that MOBA game has developed over 17 years. However, DOTA 2 and LOL are very hard to learn that makes a lot of new players jump away from the game. This situation greatly limits the development of the MOBA game. In this case, a subversive game called Heroes of Storm came out in 2015.

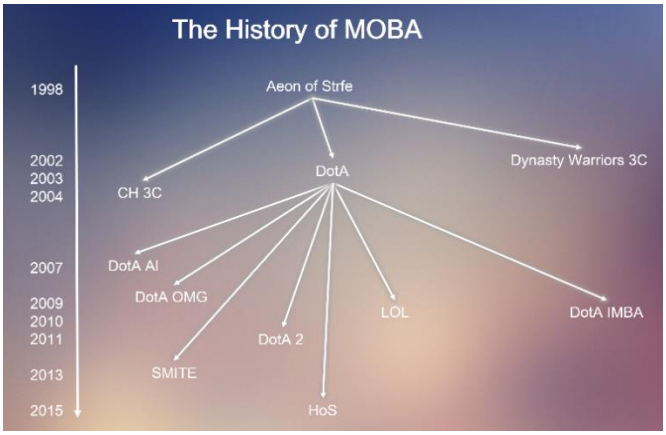


Fig 1. History of MOBA game

1.2 Research Object: Heroes of the Storm

Heroes of the Storm (originally titled Blizzard DOTA and later changed to Blizzard All-Stars) is a multiplayer online battle arena video game which has been developed by Blizzard Entertainment. The game features heroes from Blizzard's franchises including Series of Warcraft, Diablo, and StarCraft. [3].

In order to develop the game rhythm and make more interest for players, Blizzard has revised a lot of mechanisms to modify the traditional MOBA game. Heroes of the Storm revolves around online 5-versus-5 matches, operated through Blizzard's online gaming service Battle.net. Players can choose one among three game modes, which make the players play with/against computer-controlled heroes or other players. When players first start the game, they may play five heroes provided by the free hero rotation, a methodically selected list that changes weekly, but by using gold, the in-game source of wealth, or through micro transactions, they can gain permanent access to a hero. Two additional heroes are available to players who have reached level 15. As of July 2015, there are currently 39 heroes in the game divided into 4 separate roles. Of the currently released maps, 6 of the 8 have the standard 3 main lanes where players can fight, while the others have only two main lanes, but a separate objective-based area. Killing enemy/neutral units

and the opposing side's heroes grants experience points, which are shared with the entire team. When a certain experience threshold is reached for a team, each hero on that team levels up, acquiring slightly amplified status and gaining a talent point upon reaching levels 1, 4, 7, 10, 13, 16, and 20. Talent points allow players to customize their hero's abilities and generally result in large increase in power, especially for levels 10 and 20. This level-up system emphasizes the importance of teamwork, since a player's action can affect the whole team. Minions at neutral camps can be defeated to gain mercenaries that fight for the player. Each map has a different side-objective that will help either team deal significant damage to the other.

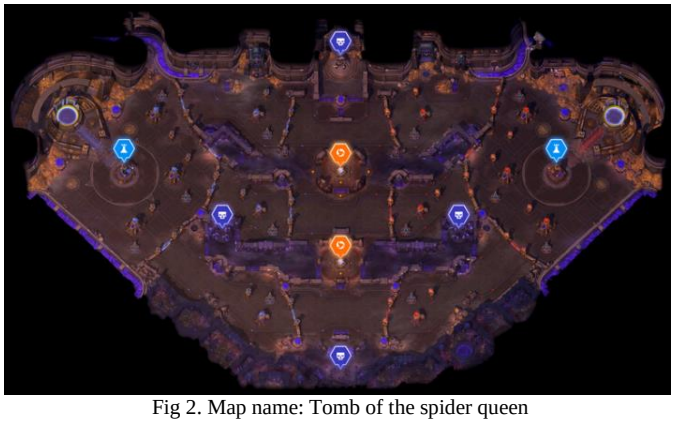


Fig 2. Map name: Tomb of the spider queen

Fig 3. Map name: Garden of Terror

There are many different maps existed in the Heroes of the Storm. For each map, players can choose the corresponding strategy to fight against each other, also the same hero can choose the different talent to fit the map environments as shown in Figure 2 and 3 [10].

2. Game Refinement Theory

2.1 Original game refinement theory

We give a short sketch of the basic idea of game refinement theory from [9]. The "game progress" is twofold. One is game speed or scoring rate, while another one is game information progress with focus on the game outcome. In sports games such as soccer and basketball, the scoring rate is calculated by two factors: (1) goal, i.e., total score and (2) time or steps to achieve the goal [6]. Thus, the game speed is given by average number of successful shoots divided by average number of shoot attempts. For other score-limited sports games such as Volleyball and Tennis in which the goal (i.e., score to win) is set in advance, the average number of total points per game may correspond to the steps to achieve the goal [10].

Game information progress presents how certain is the result of the game in a certain time or steps. Let G and T be the average number of successful shoots and the average number of shoots per game, respectively. If one knows the game information progress, for example after the game, the game progress x(t) will be given as a linear function of time t with $0 \leq t \leq T$ and $0 \leq x(t) \leq G$, as shown in Equation (1)

$$x(t) = \frac{G}{T}t \quad (1)$$

However, the game information progress given by Equation (1) is unknown during the in-game period. The presence of uncertainty during the game, often until the final moments of a game, renders exponential game progress. Hence, a realistic model of game information progress is given by Equation (2).

$$x(t) = G\left(\frac{t}{T}\right)^2 \quad (2)$$

Here n stands for a constant parameter which is given based on the perspective of an observer of the game considered. Then acceleration of game information progress is obtained by deriving Equation (2) twice. Solving it at t = T, the equation becomes

$$x''(T) = \frac{Gn(n-1)}{T^n}t^{n-2} = \frac{G}{T^2}n(n-1)$$

It is assumed in the current model that game information progress in any type of game is encoded and transported in our brains. We do not yet know about the physics of information in the brain, but it is likely that the acceleration of information progress is subject to the forces and laws of physics. Therefore, we expect that the larger the value $\frac{G}{T^2}$ is, the more the game becomes exciting, due in part to the uncertainty of game outcome. Thus, we use its root square $\frac{\sqrt{G}}{T}$, as a game refinement measure for the game under consideration. We can call it R-value for short.

2.2 Identify the Headings

Here we consider the gap between board games and sports games by deriving a formula to calculate the game information progress of board games. Let *B* and *D* be average branching factor (number of possible options) and game length (depth of whole game tree), respectively. One round in board games can be illustrated as decision tree. At each depth of the game tree, one will choose a move and the game will progress. Figure 1 illustrates one level of game tree. The distance *d*, which has been shown in Figure 1, can be found by using simple Pythagoras theorem, thus resulting in $d = \sqrt{\Delta l^2 + 1}$.

.

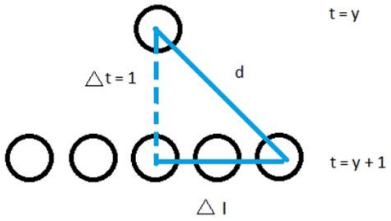


Fig.4 Illustration of one level of game tree

Assuming that the approximate value of horizontal difference between nodes is $\frac{B}{2}$, then we can make a substitution and get $d = \sqrt{(\frac{B}{2})^2 + 1}$. The game progress for one game is the total level of game tree times *d*. For the meantime, we do not consider Δt^2 because the value ($\Delta t^2 = 1$) is assumed to be much smaller compared to B. The game length will be normalized by the average game length *D*, then the game progress *x(t)* is given by $x(t) = \frac{t}{D} * d = \frac{t}{D} * \sqrt{(\frac{B}{2})^2} = \frac{Bt}{2D}$. Then, in general we have, $x(t) = c\frac{B}{D}t$, where *c* is a different constant which depends on the game considered. However, we manage to explain how to obtain the game information progress value itself. The game progress in the domain of board games forms a linear graph with the maximum value *x(t)* of *B*. Assuming $c = 1$, then we have a realistic game progress model for board games, which is given by

$$x(t) = B\left(\frac{t}{D}\right)^n \quad (3)$$

Equation (3) shows that the game progress in board games corresponds to that of sports games as shown in Equation (2).

In addition, the branching factor B can be defined as the number of possibility. For example in table tennis, each round of game could be regarded as one depth/length, for each round, there are only two possibility– win or lose. Therefore, another definition is a way to figure out the progress model of a target game using two factors: possibility result (say W) and total round of entire game (say T). R-value is given by $R = \frac{\sqrt{W}}{T}$. To support the effectiveness of proposed game refinement measures, some data of games such as Chess and Go [6] from board games and two sports games [9] are compared. We show, in Table I, a comparison of game refinement measures for various type of games. From Table I, we see that sophisticated games have a common factor (i.e., same degree of acceleration value) to feel engagement or excitement regardless of different type of games [8].

Table 1: Measures of game refinement for various game

Game	G or B	T or D	R-value
Chess	35	80	0.074
Go	250	208	0.076
Soccer	2.64	22	0.073
Basketball	36.38	82.01	0.073
DotA ver 6.48	69.2	110.8	0.075
DotA ver 6.64	68.4	110.4	0.075
DotA ver 6.80	68.6	106.2	0.078

3. Game Refinement Theory in HOS

We consider the game progress of Heroes of the Storm. It can be measured by two factors: kill heroes and destroy fortress. Let K and A be the average number of successful killing heroes and destroying fortress, and the average number of attempts per game, respectively [7]. If one knows the game information progress, for example after the game, the game progress *x(t)* will be given by Equation (4).

$$x(t) = \frac{K}{A}t \quad (4)$$

A model of Heroes of the Storm game information progress is given by Equation (5).

$$x(t) = K\left(\frac{t}{A}\right)^2 \quad (5)$$

Here n stands for a constant parameter which is given based on the perspective of an observer in the game considered. Then acceleration of game information progress is obtained by deriving Equation (5) twice. Solving it at t = A, the equation becomes

$$x''(T) = \frac{Kn(n-1)}{A^n}t^{n-2} = \frac{K}{A^2}n(n-1)$$

Therefore, the refinement value in Heroes of the Storm can be described as the Equation (6)

$$R = \frac{\sqrt{K}}{A} \quad (6)$$

As game players, the first thing they care about is how to develop themselves and limit the development of the enemy. Each player have different role in the game, therefore players have to choose the different kinds of heroes and their talent base on the corresponding map. In order to make the data more objective and reasonable, we choose the expert players' video to analyze data. The statistics was collected the data of killing and the destroyed fortress of each replay. As the Table 2 shows, the results of different map using game refinement measure by computer system, and one fortress equal to four defense force, then we have $K' = K + 4 * D$.

Table 2: Measures of game refinement for each map in Heroes of the Storm

Map	K	D	A	K'	R
Blackheart's bay	37.300	8.4	71.7	70.900	0.117
Sky temple	38.875	9.7	70.2	77.675	0.126
Dragon Shire	39.100	6.2	82.6	63.900	0.097
Tomb of the Spider queen	45.800	7.3	90.7	75.000	0.095
Infernal shrines	39.875	5.8	87.2	63.075	0.091
Cursed hollow	40.350	7.3	93.4	69.550	0.089
Battlefield of eternity	44.950	5.5	93.8	66.950	0.087
Garden of terror	37.225	7.9	81	68.825	0.102
Haunted mines	38.875	4.2	73.9	55.675	0.101

4. Discussion and Comparison

We collected data of Heroes of the Storm (HOS) for each map. Then, we applied game refinement theory application in Section II and Section III. In the previous studies, it is found that sophisticated R-value in each game between $0.07 - 0.08$. However, we see that the results show much higher values for HOS battle. It means this game will be too excited which is suitable for especial viewer such as boxing which is extremely

exciting sport. Compared with other MOBA game such as DOTA, we can summarize the conclusion as below:

- As a new game, HOS still has some insufficient aspects. In fact, until our research was done, only 43 heroes can be chosen, however, there are 112 heroes in DOTA. Many new heroes should be added and the current heroes' parameter should be changed.
- Generally, the R-value in HOS is too high even approach 0.1, and DOTA almost in the window value which between 0.07 and 0.08. It means DOTA fit to set as an e-sports competition item, but HOS is fit to do entertainment. DOTA has powerful skill and more visual impact for each hero, what cares more about management and running. Players need to make a stable and safe environment for carry and develop. Gank usually happens during the whole game. However in the HOS, the most important thing is large-scale team combat, therefore, the game rhythm is much higher than DOTA. Generally, a DOTA game may spend about 40 minutes but HOS usually within 20 minutes. HOS offers game players a new style of MOBA game that spends less time of each game and form a fast rhythm.
- According to the Table 2, the most interesting and exciting map are Sky temple and Blackheart's bay. Battlefield of eternity and Garden of terror are suitable for e-sports competition. In fact, DOTA can consider make more maps in the future to improve the fun level. The higher refinement value will be tenderer to the freshman players. DOTA is very unamiable to the new players.
- For the mechanism, DOTA focus on the anaphase period during the game, but the core mechanism in HOS is wild monster. For this reason, the game depth of HOS is less than DOTA and gets a larger R-value. Therefore, HOS cares more about teamwork not personal operation and game awareness.

Nevertheless, the fun of HOS is not derived only from battle. The various heroes and their talent can provide a lot of enjoyment for Blizzard fans. In addition, they can design maps that are more interesting. For example, control the map mechanism to keep the R-value between 0.07 and 0.08, what suitable for held e-sports use.

5. Conclusion

This paper we have extended game refinement theory to the field of HOS and builds a model of measurement for each map in HOS. The results of computer analysis confirmed HOS has the similarity of game entertainment impact like sports games and board games. It means that multi-player game also follow the principle of seesaw games. Compared with the other MOBA game such as DOTA, the game rhythm of HOS is much higher, it means player can feel more enjoyable from this game, on the other hand it shows HOS has lacked serious competition. HOS is trying to increase the amount of team play involved; this will absolutely lead to more fun than ever before. However, we need further investigation in collecting data and apply game refinement theory to another famous MOBA game such as LOL. Finally, further work may be considered the comparison of three popular MOBA games (DOTA, LOL and HOS).

References

[1] Multi player online battle arena,
http://en.wikipedia.org/wiki/Multiplayer online battle arena

[2] DotA, http://en.wikipedia.org/wiki/Dota
Heroes of the Storm, https://en.wikipedia.org/wiki/Heroes of the Storm

[3] C. Chambers, W.Feng, W.Feng, and D.Saha. (2005). Mitigating information exposure to cheaters in real-time strategy games, In Proceeding of NOSSDAV '05 Proceedings of the international workshop on Network and operating systems support for digital audio and video, pp.7–12.

[4] D.Cheng, R.Thawonmas. (2004). Case-based plan recognition for real-time strategy games. In Proceedings of the 5th Game-On International Conference, pp.36–40.

[5] H. Iida, N. Takeshita, and J. Yoshimura. (2003). A metric for entertainment of boardgames: Its implication for evolution of chess variants. Entertainment Computing Technologies and Applications, pages 65–72.

[6] H. Iida, K. Takahara, J. Nagashima, Y. Kajihara and T. Hashimoto.
(2004). An application of game-refinement theory to Mah Jong. In Entertainment Computing–ICEC2004, pp. 333–338. Springer.

[7] XIONG Shuo, ZUO Long and H. Iida (2014). Quantifying Engagement of Electronic Sports Game. Advances in Social and Behavioral Sciences,5, 37–42.

[8] A. P. Sutiono, A. Purwarianti, and H. Iida. (2014). A mathematical model of game refinement, in D. Reidsma et al. (Eds.): INTETAIN2014, LNICST 136, 148–151.

[9] J. Takeuchi, R. Ramadan, and H. Iida. (2014). Game refinement theory and its application to Volleyball, Research Report 2014-GI-31(3),Information Processing Society of Japan, 1–6.

[10] Heroes of the Storm Game guide. 2015 BLIZZARD ENTERTAINMENT, INC. ALL RIGHTS RESERVED. url: http://us.battle.net/heroes/e

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Introducing an Efficient Method for Scheduling Independent Tasks in Grid Environment using Meta-Heuristic Algorithms

Masoud Shirzadi[1], Mortaza Zolfpour-Arokhlo[2], Majid Sina[3]

[1] Department of Computer Engineering, Yasuj Branch, Islamic Azad University
Yasuj, Iran
shirzadimasoud@gmail.com

[2] Department of Computer Engineering, Sepidan Branch, Islamic Azad University
Sepidan, Iran
zolfpour@gmail.com

[3] Department of Computer Engineering, Behbahan Branch, Islamic Azad University
Behbahan, Iran
majidsina.edu@gmail.com

Abstract

Since the dynamicity and inhomogeneity of resources complicates scheduling, it is not possible to use accurate scheduling algorithms. Therefore, many studies focus on heuristic algorithms like the artificial bee colony algorithm. Since, the artificial bee colony algorithm searches the problem space locally and has a poor performance in global search; global search algorithms like genetic algorithms should also be used to overcome this drawback. This study proposes a scheduling algorithm, which is combination of the genetic and artificial bee colony algorithms for the independent scheduling problem in a computing grid. This study aims to reduce the maximum total scheduling time. Simulation results indicate that the proposed algorithm reduces the maximum execution time (makespan) by 10% in comparison to the compared methods.

Keywords: *computing grid, independent task scheduling, genetic algorithm, artificial bee colony algorithm.*

1. Introduction

The optimization methods and algorithms are divided into precise and approximate algorithms. Precise algorithms are able to find the optimum accurately; however, they are efficient for hard optimization problems and their execution time increases exponentially [1]. Approximate algorithms are able to find good (close to optimal) solutions for hard optimization problems in a short time. Approximate algorithms are also divided into heuristic, meta-heuristic, and hyper-heuristic algorithms. Two main problems of heuristic algorithms are falling into local optimums and their inability in different problems. Meta-heuristic algorithms have been proposed to overcome these issues. In fact, meta-heuristic algorithms are one of the approximate optimization algorithms with strategies to escape local optimums and they are applicable to a wide range of problems [2].

Since, grid resources are inhomogeneous; it is very difficult and complicated to develop a scheduling that can properly schedule dynamic and inhomogeneous resources. Computing grid systems proved an appropriate platform to run complex applications that require many heavy computations. In distributed systems, resources are distributed geographically in the environment; however, this distribution is transparent to users, since logically from their point of view, resources are aggregated in one spot [3].

2. Previous Works

Distributed computing systems (DCSs) are networks with high-speed interconnected processors that support parallel applications. DCSs provide a hardware architecture to run concentrated scientific computing applications. Exploiting parallel applications in DCSs dependents on the method in which tasks are scheduled on processors [2][4]. A hierarchical analysis process is used to schedule tasks and assign resources using multi-variable decision-making [1]. The genetic algorithm with a floating fitness function is used to achieve the most appropriate importance coefficients of completion time and cost of tasks for the scheduling operation [2]. In another study, a multi-parent crossover and a transformation operator are proposed instead of a mutation operator to solve optimization problems [5]. The genetic algorithm was first proposed by Holland for an optimization process [6]. Several years later, this method was complimented by Goldberg [7]. The artificial bee colony algorithm, which was proposed in [8] to optimize real parameters, is a new optimization

algorithm that simulates the searching behavior of a honeybee colony. ABC consists of three types of bees: worker bees, onlooker bees, and scout bees.
Scout bees fly over the resources that should be utilized. Onlooker bees Select resources by observing the scout bees` dance and worker bees randomly select resources using some intrinsic instincts or external signs if possible. The information exchange between honeybees is one of the most important events affecting collective knowledge formation. The dance area is the most important part of the hive regarding information exchange. Bees communicate in the dance area according to the quality of food resources. Different movements in this location, like moving, rotating, and vibrating depend on the distance from the discovered resource and its angle with the sun. This type of dance is called waggle dancing [9]. An algorithm was also proposed, which was a combination of the genetic and gravitational emulation local search (GELS) algorithms to solve independent task scheduling. Since, GA has a poor performance in local search; its combination with a GELS algorithm overcomes this drawback. The proposed algorithm focuses on two problems, time and lost tasks [10]. Moreover, the combination of artificial bee colony (ABC) and local search (LS) was used to optimize measures like average waiting time and finish time in the task-scheduling problem in the grid [11]. In contrast to single-objective algorithms, this algorithm searches the solution space comprehensively. Therefore, it is less probable to converge to local optima [12].

3. Scheduling Problem

The independent task-scheduling problem, for instance, includes N tasks and M machines. In other words, in the independent task assignment in a grid computing system, we have a number of computing resources with different processing speeds and each resource has a processing unit. The processing speed of each resource is defined in million instructions per second (MIPS). In this problem, we have a number of tasks with different instructions and we want to assign them optimally to the resources. The goal is to minimize the total execution time of the tasks assigned to each resource. The proposed algorithm only considers one of the quality of service parameters, i.e. time constraint, and ignores the cost. Each task can only be executed on one resource and it is not stopped until the end of its execution.
Since the proposed scheduling algorithm is static, it is assumed that the expected execution time of each task is predetermined on each of the resources. The total execution time of each resource is achieved separately after assigning tasks to resources. The main goal is to minimize the execution time of all tasks. In other words, we intend to minimize the single runtime of all resources. According to Eq.(1), if in possible solution (Sol) for a scheduling problem, task (i) is assigned to resource (j), the index (i) in the corresponding string is as follows:

$$Sol_i = j \tag{1}$$

The value of the cost function is obtained by Eq.(2).

$$\text{Cost} = \text{Min}\{\text{MS} = \text{Max}(\text{RT}[i,j])\} \quad 1 \le i \le N\,,\, 1 \le j \le M \tag{2}$$

Where, RT[i, j] is the execution time of task (i) on resource (j) using Eq.(3) and MS is the execution time of all input tasks (completion time).

$$\text{RT}[i,j] = \frac{JC_i}{PS_j} \qquad 1 \le i \le N, 1 \le j \le M \tag{3}$$

In equation (3), JC_i is the length or the number of instructions of tasks (i) and PS_j is the processing speed of resource (j).

4. Methodology

Since the efficiency of the genetic algorithm, highly depends on its chromosome representation, in the proposed scheduling algorithm, a simple method is used to represent chromosomes. Accordingly, natural numbers are used to encode chromosomes. In the proposed algorithm, a possible solution is represented by a chromosome. Each chromosome is a string of natural numbers with length N, where N is the total number of independent tasks in the problem. The values of the genes are random numbers between 1 to M, where M is the total number of resources.
Table (1) presents a possible solution for the scheduling problem in which N is the number of tasks assigned to 4 resources for execution. As we can see, in this solution, the second, fourth, and Nth tasks are assigned to the first resource. Moreover, the fifth tasks are assigned to the second resource, the third and seventh task to the third resource, and the first and sixth tasks to the fourth resource.

Table 1: A possible solution to the scheduling problem

T_1	T_2	T_3	T_4	T_5	T_6	T_7	…	T_N
R_4	R_1	R_3	R_1	R_2	R_4	R_3	…	R_1

In the ABC algorithm, each food resource position is a candidate solution of the scheduling problem. In other words, each bee is considered a possible solution in a formation of the problem. Each element in the artificial bee vector is a random integer between 1 to nPop, where nPop is the number of resources. The (i)th bee specifies the resource, to which a number of tasks are assigned. For

instance, table (2) shows that in the second bee vector, task 2 (T_2) runs on resource 3 (R_3).

Table 2: An example matrix of artificial bee representation

	Task 1	Task 2	Task 3	Task 4
Bee1	Resource1	Resource2	Resource3	Resource4
Bee2	Resource1	Resource3	Resource4	Resource1
Bee3	Resource3	Resource4	Resource2	Resource1

Since the bees` positions are defined in a continuous space, the proposed algorithm also defined bees` positions as continuous values. After producing a new population, the values in the position vectors may be decimal; this is invalid for the resource number. Therefore, the proposed algorithm transforms each decimal position to a valid integer. That way, a continuous optimization problem is transformed into a discrete one. Thus, the bees are evaluated in a discrete space; however, the bees` movements are in a continuous space.

In the task-scheduling problem, each solution depends on the evaluation function and, the better solution is determined based on its value. Therefore, the amount of food resources depends on the value of the evaluation functions of ABC. The number of working or onlooker bees is equal to the number of solutions in the population. ABC produces a random solution or initial population with food resource size (nPop). Each solution represents the position of a food resource, which is shown by x_{ij} and (i) indicates a specific solution (i=1, 2, … , nPop). Each solution is represented by a (D) dimensional vector and thus, (j) indicates a dimension of a certain solution (j=1, 2, … , D). After a random solution is generated, the worker bees initiate their search. Worker bees search around the previous food resource position. If the new solution is better, it replaces the last one. Food resource (solution) comparisons are based on an evaluation function (the nectar of the food resource).

After all worker bees finished the searching process, the information of food resources (solutions) and their positions are shared with onlooker bees. Now, onlooker bees select food resources based on their probability values (P_i). The probability values of food resources are computed using equation (4). Therefore, the probability value of each resource used by onlooker bees is determined after evaluating food resources by worker bees.

$$p_i=\frac{\text{fitness}_i}{\sum_{i=1}^{nPop}\text{fitness}_i} \quad (4)$$

Eq.(5) is used to generate a candidate solution from the previous solution. Therefore, the step length is increased based on the search goals and the optimal solution in the search space. After producing candidate solution (v_{ij}), its evaluation value is computed and compared with that of x_{ij}.

$$\upsilon_{ij}=x_{ij}+\phi_{ij}\left(x_{ij}-x_{kj}\right) \quad (5)$$

If the new candidate solution has an equal or more nectar (evaluation value), it replaced the previous solution. If a solution is not improved for a number of predetermined iterations, it is assumed that it has no more food resources. The finished food resource is replaced by the new one by scout bees.

5. The Proposed Scheduling Algorithm (ABC-GA)

There are five phases in this algorithm: initialization, worker bees, onlooker bees, scout bees, and mutation. We add mutation after the onlooker bee phase. The onlooker bees perform local search and mutation searches the search space and tries to find a new area of the solution space. Using the mutation operator, it is likely for the best local position to change and prevent the algorithm from falling in the local optima.

The mutation operator can be applied to the genes of the generated offspring with probability P_m (mutation coefficient). More specifically, the value of each gene in the offspring chromosome changes with probability P_m. the mutation operators are used to prevent premature convergence and falling in the local optima. In this method, two genes are randomly selected and their positions are exchanged as table (3).

Table 3: An example of the mutation operator

Initial chromosome

T1	T2	T3	T4	T5	T6	T7	T8
R3	R3	R2	R1	R2	R3	R1	R4

The mutated chromosome

T1	T2	T3	T4	T5	T6	T7	T8
R3	R3	R3	R1	R2	R2	R1	R4

For instance, if we have 100 chromosomes and each chromosome has 10 genes, mutation can be applied to some of the 1000 genes in the population. Therefore, if the mutation probability is assumed 0.05, it means that 50 genes of the 1000 gens in the population may mutate and the rest remain unchanged. If the mutation probability is one, all genes are changed and if it is zero, no change is applied to the chromosomes. We must note that P_m should not be large, since it causes diversity and genetic dispersion in the population and this dispersion significantly reduces the convergence speed of the algorithm.

In the proposed method, the probability based mutation stage is performed in each food search operator for each iteration during the life cycle of ABC`s optimization techniques. Food resources are selected randomly. In the mutation phase, the generated offspring replace the old ones. The mutation operator is an exchange operation. During mutation, food resource x_{ij} is randomly selected and one of its members is replaced with a random number between the lower and upper boundary of food resources. The stage of the proposed algorithm is as follows:

1) Initialization phase
    a. Begin
    b. Set the food resource positions (X_{ij}) randomly
    c. Calculate the evaluation function of each food resource
    d. Select the best food resource among the current resources based on the evaluation function
2) Repeat
3) Calculate the food resources` fitness values to find the best resources
4) Worker bee phase
    a. Generate a new candidate solution
    b. Calculate the evaluation value of each food resource
    c. If the evaluation value of the new candidate solution is better, replace it with the current solution.
5) Calculate the food resources` fitness values to find the best resources
6) Calculate probability (P_i) for each food resource based on the fitness value
7) Onlooker bee phase
    a. Select a food resource based on probability P_i
    b. Generate a new solution for the food resource with position X_{ij}
    c. Calculate evaluation values of food resources
    d. If the evaluation value of the new candidate solution is better, replace it with the old solution
8) Mutation phase
    a. If the mutation condition is satisfied,
        i. Select a random member of the current population for mutation
        ii. Apply mutation to produce new food positions
9) Scout bee phase
    a. If any of the food resources are finished
        i. Replace it by the positions randomly generated by the scout bee
        ii. Calculate evolution values for the new positions
        iii. Remember the best solution so far
10) Repeat until stop condition is met

6. Results

All experiments were conducted on a system with a 2.40 GHz processor, 4G RAM, and Windows 7. The Matlab environment is used for the simulation of all algorithms. We must note that each algorithm is evaluated with different parameters and operators several times and finally, the best values are selected for the parameters and the best operators are selected for each algorithm. In fact, each algorithm is simulated under the best conditions.
Before discussing the results, we should specify the initial values of the parameters in the proposed algorithm ABC-GA. Tables (4) and (5) present the initial parameter values of each algorithm.

Table 4: The initial value of the mutation parameter in the genetic algorithm

Parameter	Mutation coefficient
Value	0.004

Table 5: The initial parameter values in the artificial bee colony algorithm

Parameter	No. Bees	No. Food resources
Value	50	50

For the first experiments, the proposed algorithm was compared with several other scheduling algorithms, according to the conditions in table (6). All models were considered equal to properly investigate the tasks` lengths in the experiments. In other words, this parameter considers the number of instructions in each task in a uniform distribution range. The number of iterations parameter indicates that 200 iterations are performed to obtain the execution time of the program using the existing algorithms and the mean value is selected for evaluations.

Table 6: The experimental conditions of time optimization with variable number of tasks

Parameter	Algorithm	No. Users	No. Tasks	Task Lengths	No. Iterations
Value	variable	1	variable	[10…30]	200

These conditions are considered to evaluate different algorithms. Under the conditions in table (6), the proposed algorithm was compared with the artificial bee colony algorithms, genetic algorithm, particle swarm optimization, and firefly algorithm.

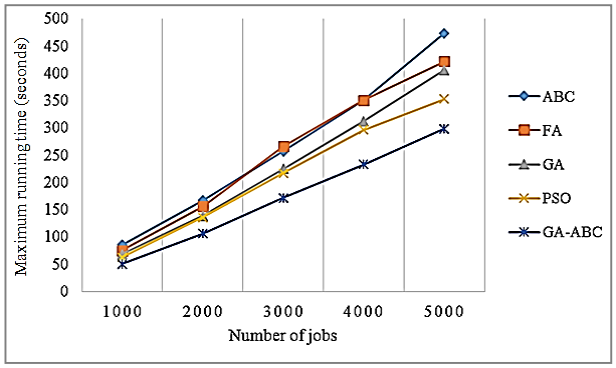


Figure 1. The completion time with different requests.

As we can see, the more input tasks there are, the completion time is longer. Among the scheduling algorithms, the proposed algorithm (ABC-GA) has a shorter completion time. This experiment shows that a larger number of tasks increase the time difference between the completion time of the proposed algorithm and that of other methods.

The second experiment compares the scheduling algorithms with a time optimization strategy and tasks with different heterogeneities. The proposed algorithm is compared with other methods under the conditions specified in table (7). The task length range parameter is variable and experiments are conducted in different ranges (different heterogeneities). The number of iterations shows that 200 iterations are performed to obtain the algorithms` runtimes for certain heterogeneity and the mean of the values is selected for evaluation.

Table 7: The experimental conditions of time optimization with a variable task length range

Parameter	Algorithm	No. Users	No. Tasks	Task lengths	No. Iterations
Value	variable	1	5000	variable	200

As we can see in figure (2), the time changes due to increasing the heterogeneity of tasks are different for each algorithm. Increasing the heterogeneity of task lengths also increases the runtime of the algorithms. According to figure (2), the proposed algorithm has a shorter completion time in comparison to other methods.

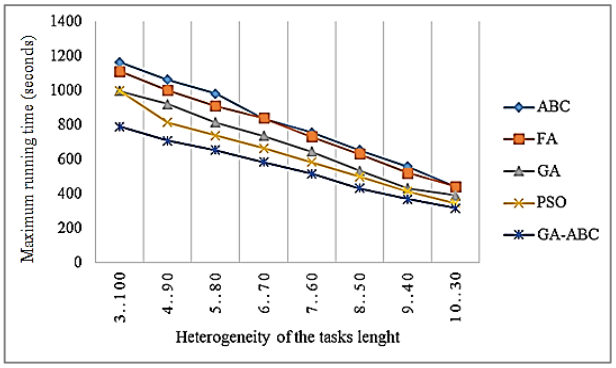


Figure 2. Comparison of different algorithms` runtimes for different task length heterogeneities.

The last experiment compares the proposed algorithm and several other scheduling methods under the conditions of table (8). All models are considered equal to properly investigate the number of resources. The number of iterations shows that 200 iterations are performed to obtain the algorithms` runtimes for certain heterogeneity and the mean of the values is selected for evaluation.

Table 8: Experimental conditions of time optimization with different numbers of resources

Parameter	Algorithm	No. Users	No. Tasks	Task lengths	No. Resources	No. Iterations
Value	variable	1	5000	[10…30]	variable	200

The conditions above are used to evaluate different algorithms. Under such conditions, the proposed algorithm is compared with the artificial bee colony, genetic algorithm, particle swarm optimization, and firefly algorithm.

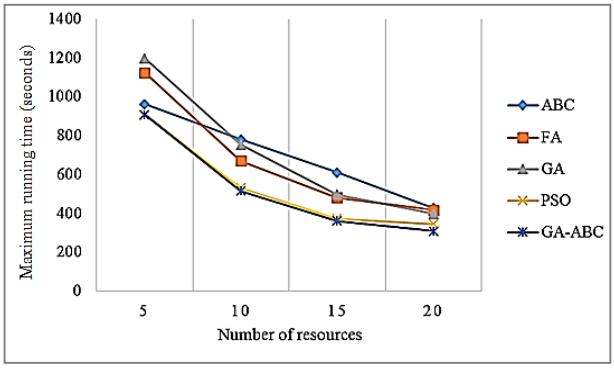


Figure 3. The completion times with different numbers of resources.

In figure (3), the effect of different numbers of resources on the maximum execution time of tasks is presented for the proposed algorithm and the compared methods. The number of resources is considered variable. As we can see, time changes due to increasing the number of resources are

different for each algorithm. Increasing the number of resources reduces the runtime of the algorithm. As we can see, the proposed algorithm has a shorter completion time in comparison to other methods.

7. Conclusions

This research presented a meta-heuristic combination of the artificial bee colony (ABC) and the genetic algorithm (GA) for the scheduling problem. The ABC algorithm is one of the good heuristic algorithms due to features like fault tolerance, flexibility, independence from initial values, and high convergence speed. Combining the features of ABC with those of GA can accelerate the convergence and the identification of the optimal solution. The proposed algorithm was compared with a number of well-known optimization algorithms and results indicated that it has a shorter completion time.

References

[1] Ahmadi Mahmoodabadi, A., Mehri Tokmeh, J., and Habibi Zadnovin, A. 2013. A task-scheduling algorithm by multi-criteria decision-making in grid environment. 10th National Conference on Computer and Intelligent Systems, 7-1.

[2] Ashrafkia, S., Mirnia, M.K., and Habibi Zadnovin, A., 2013. Scheduling in grid environment using genetic algorithms with floating fitness function. 10th National Conference on Computer and Intelligent Systems, 6-1.

[3] Yaghini, M. and Akhavan Kazemzadeh, M.R., 2014. Meta-heuristic optimization algorithms. Tehran: Jihad Daneshgahi Publication (Amirkabir University of Technology).

[4] Daoud, M.I. & Kharma, N. (2011). A hybrid heuristic–genetic algorithm for task scheduling in heterogeneous processor networks, J Parallel Distrib Comput.71:1518-1531.

[5] Elsayed, S.M., Sarker, R.A. & Essam, D.L. (2014). A new genetic algorithm for solving optimization problems, Engineering Applications of Artificial Intelligence. 27, 57-69.

[6] Holland, J. (1975). Adaptation in Natural and Artificial Systems, MIT Press Cambridge, ISBN: 0262581116, p.228.

[7] Goldberg, D. E. (1989). Genetic Algorithms in Search Optimization and Machine Learning, Addison-Wesley, ISBN: 0201157675, p.432.

[8] Karaboga, D. (Oct 2005). An idea based on honeybee swarm for numerical optimization.

[9] Akay, B. & Karaboga, D. (2012). A modified Artificial Bee Colony algorithm for real-parameter optimization. Information Sciences. 192, 142-120.

[10] Pooranian, Z., Shojafar, M., Abawaji, J.H. & Singhal, M. (2013). GLOA: A New Job Scheduling Algorithm for Grid computing. International Journal of Artificial Intelligence and Interactive Multimedia, 2(1), 59-64.

[11] Tammano, A. & Phu-ang, A. (2013). A Hybrid Artificial Bee Colony Algorithm with Local Search for Flexible Job-Shop Scheduling Problem. Procedia Computer Science. 20, 96-101.

[12] Parvan, H., Behrouzian-Nejad, E. & Alavi, S.E. (2014). Tasks Scheduling in Computational Grid Based on Meta-Heuristic Algorithms, International journal of Computer Science & Network Solutions. 2, 48-54.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Reverse Modeling and Autonomous Extrapolation of RF Threats

Sanguk Noh[1] and So Ryoung Park[2]

[1] School of Computer Science and Information Engineering, The Catholic University of Korea
Bucheon, 420-743, Republic of Korea
sunoh@catholic.ac.kr

[2] School of Information, Communications, and Electronics Engineering, The Catholic University of Korea
Bucheon, 420-743, Republic of Korea
srpark@catholic.ac.kr

Abstract

This paper addresses the investigation of the basic components of reverse modeling and autonomous extrapolation of radio frequency (RF) threats in electronic warfare settings. To design and test our system, we first model RF threats using the radioactive parameters received. The enemy radar simulated with a transponder or emitter transmits electronic signals; next, the sensors of the system intercept those signals as radioactive parameters. We generate the attributes of RF threats during communication between the electronic emissions of RF threats and the receivers of our system in various electronic warfare scenarios. We then utilize the data acquired through our system to reversely model RF threats. Our system carries out the reverse extrapolation process for the purpose of identifying and classifying threats by using profiles compiled through a series of machine learning algorithms, i.e., naive Bayesian classifier, decision tree, and k-means clustering algorithms. This compilation technique, which is based upon the inductive threat model, could be used to analyze and predict what a real-time threat is. We summarize empirical results that demonstrate our system capabilities of reversely modeling and autonomously extrapolating RF threats in simulated electronic warfare settings.

Keywords: *Autonomous reverse extrapolation of threats; Data Mining using machine learning algorithms; Modeling and generating attributes of threats; Simulated electronic warfare settings.*

1. Introduction

Despite of potential danger in electronic warfare (EW) environments, first of all, our agents need to reversely extrapolate and autonomously identify threats in order to ensure their continual functionality. This paper investigates the basic components of reverse modeling and autonomous extrapolation of radio frequency (RF) threats in simulated EW settings. Autonomous situation awareness includes that the sensors perceive the signals of a dynamically changing environment, and the agents accumulate the processed data into knowledge bases. The critical step is to make the use of a specific knowledge to predict what kinds of situation will happen in an imminent future. The agents can be equipped with tracing and recognizing the state of incessantly changing and urgent environments. It is not a simply uncalculated response to a given snapshot but an elevated intelligence to make the agents adaptively operate. Thus, autonomous situation awareness is an indispensable component for an agent to be rational in the process of formulate its adaptive knowledge. This function can be widely applied to various fields such as battleground situation, traffic situation, and any kind of disaster situation [1, 2, 3].

For the reverse modeling and extrapolation of RF threats, we are obliged to use the observed or estimated attributes in place of the real attributes. In electromagnetic transmission, the radiated signal from transmitter will be modified and distorted for several reasons, and then arrived at the receiver [4, 5, 6, 7, 8, 9]. The signal power will be modified by the atmospheric loss, antenna gains, hardware losses, weather condition, and so on. The signal frequency will be transformed by the relative velocity of the RF threat and receiver. Under multipath fading environment, the signal may spread with some delay spread factor. That is, the observed attributes at the receiver for the reverse modeling could be considerably different form the real attributes at the transmitter in RF threats. To generate the observed attributes for the reverse modeling and to estimate the real attributes from the observed one, we examine the modifying principle of the electromagnetic waves during transmission in battlefield scenarios.

Given observed attributes of RF threats sensored by our agents in electronic warfare settings, we suggest a reverse extrapolation mechanism of RF threats through machine learning algorithms, i.e., both supervised naive Bayesian classifier [10] inductive decision tree algorithm [11], and unsupervised k-means clustering algorithms [12]. For our agents to have a reverse model of RF threats in a specific situation, we endow them with a set of operational knowledge. The knowledge formulated is constructed by compiling threat systems and their attributes into the resulting outputs of three machine learning algorithms. The compiled knowledge accumulated offline can be obtained from both supervised and unsupervised machine

learning algorithms. In this paper, further, the performance of each compilation method is measured so as to compare its accuracy with the others. The various compilations provide our agents with a spectrum of approaches to extrapolating reverse models under dangerous situations in EW settings.

To differentiate the types of RF threats, for example, search radars, tracking radars, and missile guidance seekers, we abstract reverse models from several types of threats in the simulated EW settings using compilation techniques. Applying both supervised and unsupervised machine learning algorithms to finding regularities has been used to detect specific patterns in many domains [13, 14] but, to our best knowledge, it could be one of new attempts for the reverse extrapolation of RF threats in electronic warfare scenarios. In our framework, both of the supervised and unsupervised machine learning algorithms compile the example situations into an operational knowledge to be applicable for autonomous situation awareness. Our approach leads to reversely model RF threats, to recognize given situation at hand based upon the compiled model, to rapidly respond to the fatal condition, and, as a consequence, to enhance our agents' continual survival.

The following section addresses the representative attributes of RF threats and design our rational agents which are equipped with reverse models extrapolating RF threats. We further generate the attributes of RF threats that realistically simulate electronic warfare scenarios given any RF threat. Section 3 describes our agent's reverse extrapolation process of threat identification in detail. Section 4 evaluates our framework empirically, and analyzes the experimental results. In conclusion, we summarize our result and discuss further research issues.

2. Analyzing Reverse Models of RF Threats and Generating Attributes for Reverse Modeling

To reversely extrapolate threats given in electronic warfare settings, we first abstract features from various RF threats and then model RF threats using the radioactive parameters received. In this section, we formulate the electronic signals of the RF threats into possible parameters for their simulated reverse extrapolation and design the architecture of our agents being capable of processing the reverse extrapolation.

2.1 Reversely Modeling RF Threats and Designing Reverse Extrapolation Process

Since our agents are assumed to perceive a threatening situation only through their radar receivers in EW settings, the RF threats that they can detect are divided into search radar, tracking radar, and missile guidance seeker [15, 16]. The RF threats can be applied to land-based, shipborne, and airborne radar systems based on the platform. Before we implement all the platforms, as the first step, we will test our agents which can be operational on the land-based platform [5, 15].

The representative attributes for agents' reverse model of RF threats in EW settings are described in Table 1. The signals perceived by radar receivers are translated into a set of variables. Given the variables, the attributes that can characterize the threats should be picked up. The attributes in Table 1 are determined to effectively discriminate three threat types among all potential threats. As shown in Table 1, the attributes acquired from radar sensors are radar frequency, pulse width, pulse power, and pulse repetition interval (PRI). The second column of Table 1 presents their values in specific ranges, and the third column describes three threat types identified, i.e., search radar, tracking radar, and missile guidance seeker.

Table 1: Relevant attributes modeling RF threats and threat types

Attributes	***Ranges***	***Threat Types***
Radar Frequency	3MHz ~ 40GHz	Search Radar / Tracking Radar / Missile
Pulse Width	0.1 ~ 5 μs	
Pulse Power	1KW ~ 1MW	
Pulse Repetition Interval (PRI)	1 μs – 1 ms	

The final goal in this research is to design and develop autonomous agents that can reversely extrapolate RF threats represented by the above attributes in Table 1, while operating in simulated electronic warfare settings. The reverse extrapolation will be extended to the range that our agents can identify not only threat types but also their block diagrams. We plan to create the block diagram which presents a certain operational principle of each threat, such as track-while-scan (TWS) radar or continuous-wave (CW) radar. In this paper, the first step towards this end is to acquire the characteristic signals of threats, and to reversely extrapolate the threat systems. The enemy's radar system simulated with a transponder or emitter transmits electronic signals; next, the radar sensors of our agents receive those signals as radioactive parameters. Given raw data sensed, the preprocessing module of our system further extracts more radioactive

variables. We then reversely extrapolate the threats into one of search radar, tracking radar, and missile guidance seeker based upon categories compiled during off-line. The architecture of reverse extrapolation system is illustrated in Fig. 1.

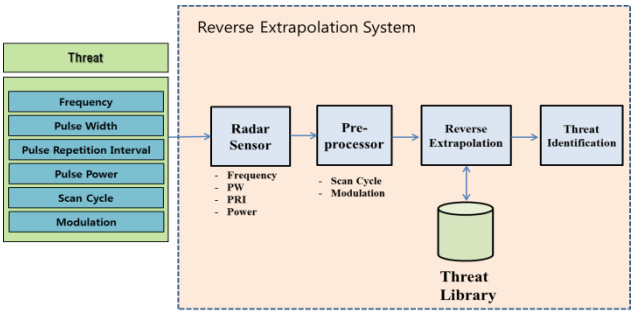


Fig. 1 The architecture of reverse extrapolation system

2.2 Generating Attributes for Reverse Modeling

To obtain the observed attributes of an RF threat at the receiver in electronic warfare settings, we examine the modifying principle of the electromagnetic waves during transmission and the estimating method of a real attribute from the modified one.

2.2.1 Modification of Signal Power

The most significant loss in power is the free-space path loss which is proportional to the square of the distance between the transmitter and receiver, and also proportional to the square of the frequency of the electromagnetic wave. In the far field where spherical spreading can be assumed, the free-space path loss L_{fs} can be expressed as [6]

$$L_{fs} = \left(\frac{4\pi f_0 R}{c}\right)^2, \qquad (1)$$

where f_0 is the frequency of RF signal, R is the distance between RF threat and receive, and c is the velocity of light. Other losses generated from various environments can be considered. The losses by atmospheric absorption due to oxygen L_{oxy} and water vapor L_{wv} are given by the Van Vleck equations [8, 9]. The loss due to rain L_{rain} increases with increased rainfall rate and radar frequency [4]. The losses L_{HW} generated by hardware (operator, collapsing, filter mismatch, and so on) may be considered if necessary [4].

Using one-way attenuation model, the received power P_r can be calculated from the transmitted power P_t by

$$P_r = \frac{P_t G_t G_r}{L_{total}} \qquad (2)$$

where $L_{total} = L_{fs} L_{oxy} L_{wv} L_{rain} L_{HW}$, G_t and G_r the transmitted and received antenna gain, respectively. After sensing the RF signal at the receiver, the received power and frequency can be observed. Then, we can estimate the transmitted power of RF signal as

$$\hat{P}_t = \frac{P_r \hat{L}_{total}}{G_r^2}. \qquad (3)$$

We assume that the transmitted antenna gain be equal to the received antenna gain and $\hat{L}_{total}$ is the calculated with the estimated distance and frequency considering the atmosphere and weather conditions.

The first row in Table 2 shows an example of the modification of RF signal power. When $P_t = 50$kW, $R = 30$km, $f_0 = 10$GHz, $G_t = G_r = 20$dB, and rainfall rate is 12.5mm/h, the received power is about 1.16W and the estimated power is about 450kW assuming that the estimated distance is 32km.

2.2.2 Modification of Signal Width

When a pulse is passed through a high-pass filter, the result is a positive spike at the leading edge and a negative spike at the trailing edge. By using the positive spike to start a counter and the negative spike to stop the count, it is possible to very accurately measure the pulse width [4]. However, under the multipath fading environment, the pulses from multipath do not arrive at the same time since the path lengths are different from each other. Then, a pulse will spread and consequently the pulse width will widen. In general, delay spread can be interpreted as the difference between the time of arrival of the earliest significant multipath component (typically, the line-of-sight component) and that of the latest components. Denoting the power delay profile of the multipath channel by $A_c(\rho)$, the mean delay of the channel is [7]

$$\bar{\rho} = \frac{\int_0^\infty \rho A_c(\rho) d\rho}{\int_0^\infty A_c(\rho) d\rho}, \qquad (4)$$

and the root mean square (rms) delay spread is given by

$$\rho_{rms} = \sqrt{\frac{\int_0^\infty (\rho - \bar{\rho})^2 A_c(\rho) d\rho}{\int_0^\infty A_c(\rho) d\rho}}. \qquad (5)$$

When a pulse with width τ is transmitted through the multipath fading channel with rms delay spread ρ_{rms}, the

observed pulse width at the receiver can be expressed as $\hat{\tau} = \tau + \rho_{rms}$. The second row in Table 2 shows an example of the modification of RF signal width. When $\tau = 0.5\,\mu s$ and $\rho_{rms} = 0.07\,\mu s$, the observed pulse width is about $0.57\mu s$. If the rms delay spread measures $0.05\mu s$, the pulse width will be estimated at $0.52\mu s$.

2.2.3 Modification of Signal Frequency

When the transmitter or receiver is moving, a change in frequency of electromagnetic waves, namely Doppler shift can be occurred. Generally, the observed frequency at the receiver f_r is given by [6]

$$f_r = \left(\frac{c + v_r}{c + v_t}\right) f_0 \approx \left(1 + \frac{\Delta v}{c}\right) f_0, \qquad (6)$$

where f_0 is the emitted frequency at RF threat, c is the velocity of light, v_r is the velocity of receiver, v_t is the velocity of RF threat, and Δv is the velocity of the receiver relative to RF threat. When the observed frequency at the receiver is f_r, the estimated frequency can be expressed as

$$\hat{f_0} = f_r / \left(1 + \frac{v_r}{c}\right), \qquad (7)$$

assuming that the velocity of RF threat is unknown and setting zero. The third row in Table 2 shows an example of the modification of RF signal frequency. When $R = 30$km, $f_0 = 10$GHz, $v_r = 290$m/s, and $v_t = 10$m/s, the observed frequency is to be nearly 10GHz.

Table 2: An example of attributes in the case of $\boldsymbol{R = 30}$km

Attributes	***Ranges***	***Threat Types***	***Estimated Values***
Radar Frequency	3MHz ~ 40GHz	Search Radar / Tracking Radar/ Missile	453kW
Pulse Width	0.1 ~ 5 μs		0.52 μs
Pulse Power	1KW ~ 1MW		10GHz

3. Reverse extrapolation of RF threats

To make our agents adaptable to simulated EW settings, we use machine learning algorithms, i.e., naive Bayesian classifier, inductive decision tree algorithm, and *k*-means clustering algorithm, and compile the example scenarios of RF threats into the resulting model of output.

As a supervised machine learning algorithm, in this section, we consider a naive Bayesian classifier and an inductive decision tree algorithm. A naive Bayesian classifier in simulated EW settings can be defined as follows: [2].

$$P(h_j \mid x_i) = \frac{P(x_i \mid h_j)P(h_j)}{\sum_{j=1}^{m} P(x_i \mid h_j)P(h_j)} \qquad (8)$$

where

- a set of attributes of an RF threat, $X = \{x_1, x_2, \ldots, x_n\}$;
- a set of types (or classes) of an RF threat, $H = \{h_1, h_2, \ldots, h_m\}$;
- $P(h_j|x_i)$ is the posterior probability of types of an RF threat h_j, $h_j \in H$, given that x_i, $x_i \in X$, is an observable attribute of an RF threat.

In our electronic warfare environments, the set of attributes of an RF threat, X, includes those described in Table 1, and the set of types of an RF threat is composed of search radar, tracking radar, and missile guidance seeker. Given a set of training data in this domain, Bayes theorem allows our agents to assign the posterior probabilities of types of an RF threat, $P(h_j|x_i)$. Our agents calculate $P(h_j|x_i)$ during online, and determine the specific type of an RF threat as the probability of a specific threat is greater than those of the others.

The decision tree approach such as ID3, C4.5 [11] and CN2 [17] uses a strategy of divide-and-conquer, which partitions the whole domain space into several types of an RF threat $C = \{c_1, c_2, \ldots, c_m\}$. From other point of view, the inductive decision tree algorithm is to find out a set of ordered attributes of an RF threat, $X = \{x_1, x_2, \ldots, x_n\}$, which separates the RF threats into a correct model with the highest information gain first. A decision tree has internal nodes labeled with attributes of an RF threat $x_i \in X$, arcs associated with their parent attributes, and leaf nodes corresponding to a set of types of an RF threat $c_j \in C$. We thus generate a decision tree representing the reverse model of various RF threats to our agents in the simulated EW setting. Based upon the generated tree, the output model can be obtained and used to interpret a new threat environment for the purpose of deciding whether any potential threat is encountered or not.

As an unsupervised machine learning algorithm, we also consider a k-means clustering algorithm [12] that aims at converging to a local optimum in an iterative refinement fashion. Given a set of instances or examples, $\{y_1, y_2, \ldots, y_n\}$, where an instance is a m-dimensional vector of attributes, the algorithm is to partition n instances into the k sets of $S=\{S_1, S_2, \ldots, S_k\}$ so as to minimize V in the following equation (9)

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |y_j - \mu_i|^2 \quad (9)$$

where μ_i is the mean of instances in S_i.

To measure the distance between two instances in the k-means clustering framework, we deploy two metrics, i.e., Euclidean distance and cosine similarity. The Euclidean distance from *a* to *b* is given by

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_m - b_m)^2}, \quad (10)$$

where $a=(a_1, a_2, \ldots, a_m)$ and $b=(b_1, b_2, \ldots, b_m)$ are two instances in Euclidean m-space. In a similar way, the distance between two instances $a=(a_1, a_2, \ldots, a_m)$ and $b=(b_1, b_2, \ldots, b_m)$ using cosine similarity is given by

$$d(a,b) = \frac{a \cdot b}{\|a\| \cdot \|b\|} = \frac{\sum_{i=1}^{m} a_i \times b_i}{\sqrt{\sum_{i=1}^{m} a_i^2} \times \sqrt{\sum_{i=1}^{m} b_i^2}}. \quad (11)$$

We thus utilize both supervised and unsupervised machine learning algorithms mentioned above to inspire our agents with a reverse model of RF threats. Our agents equipped with the resulting models generated during offline are able to reactively cope with online situation. Given an electronic warfare state, our agents apply the best reverse model among compiled models to the state, and then realize what type of RF threats is given. In this line of approach [2], the offline computation for a set of compilation significantly reduce the response time and provide our agents with more chance to survive while having more time to react.

4. Experimental Result

To evaluate the performance of reverse extrapolation process for threat identification, we generate the simulation data using discrete uniform distribution and test the compiled models by applying them to simulated electronic warfare (EW) settings. For this experiment, we use WEKA (Waikato Environment for Knowledge Analysis) [18] for supervised machine learning algorithms, i.e., naive Bayesian classifier and decision tree algorithm, and implement k-means clustering algorithms as an unsupervised technique using Euclidean distance and cosine similarity metrics, respectively. We measure the performance of our agents with reverse models in terms of the correct identification of RF threats.

4.1 Compiled Models of RF Threats

In our experiment, we applied the theoretical background of Section 2 to our simulated EW settings for the generation of attribute values. For supervised machine learning algorithms, the training data consisted of a set of attributes, i.e., radar frequency, pulse width, pulse power, and pulse repetition interval (PRI), and a class, i.e., search radar, tracking radar, and missile guidance seeker, as specified in Table 1. For unsupervised machine learning algorithms, the training data were composed of only a set of attributes without an assigned class. In our experiment, the number of total instances for training was 3, 000.

To endow our agent with three reverse models of threat data, then, the threats as training data were compiled into a set of outputs, i.e., a statistical model, an inductive rule, and a number of clusters. For the naive Bayesian classifier, the resulting output was presented as a statistical model specifying the probability of occurrence of each attribute value given a class of RF threats. C4.5 as a decision tree algorithm presented its output as a set of reactive rules. The trained result of k-means clustering algorithm was a distribution of clusters mapping from the attributes of threats to the types of RF threats.

An example of statistical model compiled through the naive Bayesian classifier was described in Table 3. Since all of attributes were numerical or continuous, in our domain, its compiled output model was the mean and the variance of attribute values. We then calculated the probability distribution of the output values given a class using normal Gaussian distribution.

Table 3: An example of statistical model compiled through naïve Bayesian classifier

Attributes	***Classes***		
	Search Radar	***Tracking Radar***	***Missile***
Radar Frequency (GHz)	1.92 ± 1.14	6.07 ± 1.18	23.56 ± 9.51
Pulse Width (μs)	3.23 ± 1.04	1.16 ± 0.21	0.45 ± 0.19
PRI (μs)	504.50 ± 307.67	3.53 ± 0.79	2.05± 0.56
Pulse Power (KW)	280.32 ± 126.98	55.89 ± 25.98	26.51 ± 13.91

The output model of reactive rules compiled by C4.5 was described in Table 4. Based on the resulting model of a decision tree, one of compiled rules was "if (pulse_width > 0.79) and (pulse_width ≤ 1.50), then tracking_radar."

Table 4: An example of rules compiled through C4.5 decision tree

Classes	***Rules***
Search Radar	**if** (Pulse_Width > 0.79) **and if** (Pulse_Width > 1.50), **then** search_radar.
Tracking Radar	**if** (Pulse_Width > 0.79) **and if** (Pulse_Width ≤ 1.50), **then** tracking_radar.
Missile	**if** (Pulse_Width ≤ 0.79), **then** missile.

Table 5 and Table 6 indicated the outputs compiled by k-means clustering algorithm using the metrics of Euclidean distance and cosine similarity, respectively. The attributes values were normalized from 0 to 100 and, from each attribute perspective, the resulting values denoted the centers for each cluster, which referred to the means nearest to a prototype of the cluster. For example, in Table 5, the 933 instances of 'search radar' belonged to the cluster 1, and the 67 instances of the same class belonged to the cluster 2. Since the cluster 1 consisted of only 'search radar,' thus, the cluster 1 should be classified into the class of 'search radar' as a result. For autonomous situation awareness, the three resulting models of RF threats could widely be used in various EW situations.

Table 5: An example of cluster compiled through K-means clustering algorithm using Euclidean distance metric

Attributes	***Clusters***		
	Cluster 1	***Cluster 2***	***Cluster 3***
Radar Frequency (GHz)	4.87	71.13	17.94
Pulse Width (μs)	66.92	6.95	19.33
PRI (μs)	53.29	0.10	1.06
Pulse Power (KW)	57.46	4.99	10.33
Attributes	***Clusters***		
	Cluster 1	***Cluster 2***	***Cluster 3***
Search Radar	***933***	67	0
Tracking Radar	0	***1000***	0
Missile	0	300	***700***

Table 6: An example of cluster compiled through K-means clustering algorithm using Cosine similarity metric

Attributes	***Clusters***		
	Cluster 1	***Cluster 2***	***Cluster 3***
Radar Frequency (GHz)	4.81	59.03	13.53
Pulse Width (μs)	63.16	7.11	29.23
PRI (μs)	58.68	0.10	2.08
Pulse Power (KW)	56.94	4.87	13.32
Attributes	***Clusters***		
	Cluster 1	***Cluster 2***	***Cluster 3***
Search Radar	***830***	0	170
Tracking Radar	0	1	***999***
Missile	0	***999***	1

4.2 Performance of Compiled Models

First, we need to find a meaningful size of the training set which could guarantee the soundness of the learning hypothesis compiled by supervised machine learning algorithms including naive Bayesian classifier and C4.5 decision tree algorithm. We set up a bunch of training examples using discrete uniform distributions starting with 180 instances. In this learning curve, we found that the sufficient number of training instances was 480, as circled in Fig. 2. The learning curves show the resulting performances (%) vs. the sizes of training examples for three RF threat types, as depicted in Fig. 2.

The naive Bayesian classifier quickly acquired the reverse extrapolation process of RF threats, as shown in Fig. 2. Its best performance turned out 100% of correctness, while those of C4.5 decision tree algorithm did 99.60%, which was almost same as the best performance. In our simulated EW settings, the performance obtained by naive Bayes classifier was a little better than that of C4.5 decision tree algorithm

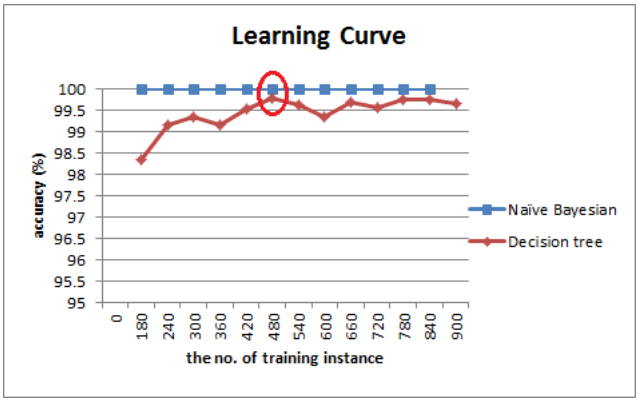


Fig. 2 The resulting performances (%) vs. the training data size for three RF threat types.

The output models compiled using supervised learning algorithms were tested by newly generated ten sets of 480 instances, which was optimally determined in Fig. 2. We could obtain the performances of the reverse extrapolation methods, as described in Table 7. Regarding the performance of the *k*-means clustering algorithm as a unsupervised learning technique, the ten sets of 3,000 instances divided into three (= *k*) classes of 1,000 ones were generated with three different initial centroids (means).

Table 7: Performances of compilation methods

Compilation Methods		***Performances***
Naïve Bayes		99.92 ± 0.11
C4.5		99.60 ± 0.11
ANOVA		$f = 46.75$
Compilation Methods	***Distance Metric***	***Performances***
K-means Clustering	Euclidean Distance	85.63 ± 2.27
	Cosine Similarity	93.50 ± 0.53
ANOVA		$f = 114.22$

We analyze the performance results in Table 7 using the standard analysis of variance (ANOVA) method. Since the computed values of $f = 46.75$ and $f = 114.22$ in ANOVA exceed 8.29 (= f .01,1,18) from the F distribution, respectively, we know that the performance of our agents, controlled by naive Bayesian classifier and C4.5 decision tree algorithm, shows meaningful difference in EW situations. In other words, the difference in their performance is not due to chance with probability of 0.99. Likewise, the performance between Euclidean distance and cosine similarity metric in case of k-means clustering algorithm reveals the same result with the above. In Table 7, the average performance of our agent using the naive Bayesian classifier in a simulated EW situation is slightly better than that of C4.5, while the agent using k-means clustering algorithm with cosine similarity metric outperforms the other agents with Euclidean distance metric.

4.3 Implementation of Test Programs

For a reverse extrapolation system equipped with outputs compiled through three machine learning algorithms, we separately implemented test programs using C# programming language. The reverse extrapolation of RF threats using naive Bayesian classifier is depicted in Fig. 3. To test the compiled knowledge, users input radioactive parameters for each attribute of RF threats at the left side of Fig. 3, select a specific algorithm, and then press the 'execute' button. The result of extrapolation is displayed at the bottom of left side, and the output panel, that is, the right side of Fig. 3 shows a statistical model and the result of threat identification given specific input parameters, as highlighted in red color.

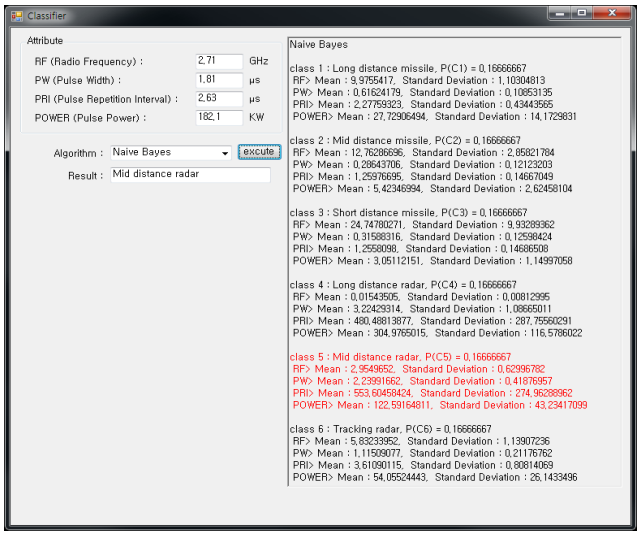


Fig. 3 The resulting reverse extrapolation using naïve Bayesian classifier.

Similarly, Fig. 4 and Fig. 5 show the reverse extrapolation using C4.5 inductive decision tree algorithm in the forms of tree diagram and text mode, respectively. The reverse extrapolation using k-means clustering algorithm, as depicted in Fig. 6, consists of four vertical axes representing four attributes, i.e., radio frequency, pulse width, pulse repetition interval (PRI), and pulse power, six horizontal lines for six classes in detail, and one resulting horizontal line as an output class. In Fig. 6, another

horizontal line of violet color comes up on the screen indicating that the resulting extrapolation class is 'an early warning search radar.'

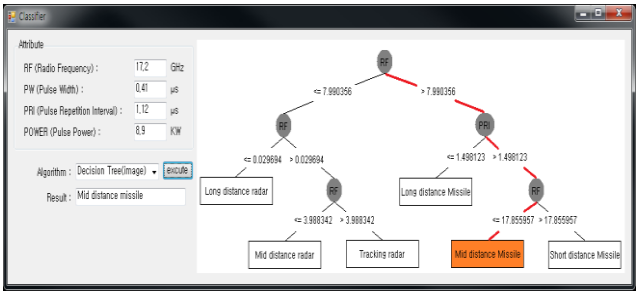


Fig. 4 The resulting reverse extrapolation using C4.5 decision tree diagram.

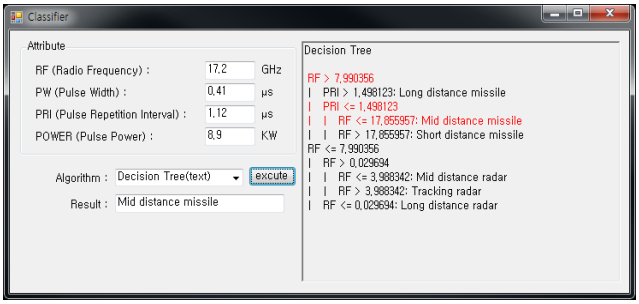


Fig. 5 The resulting reverse extrapolation using C4.5 decision tree in text mode.

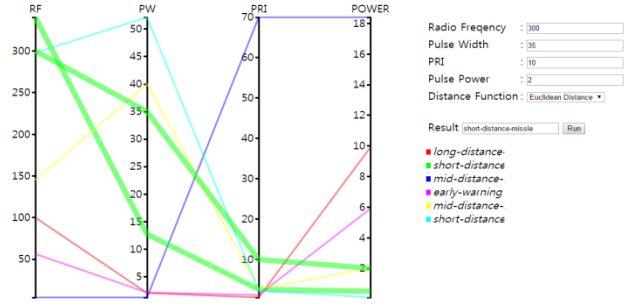


Fig. 6 The resulting reverse extrapolation using K-means clustering algorithm.

5. Conclusion and Future Work

It is indispensable for our agents to be equipped with capabilities of detecting threat signals, analyzing an electromagnetic environment, and providing a fast precise assessment of RF threats in simulated EW settings. In this paper, we showed a fully autonomous agent that reversely extrapolates various types of RF threats by using compilation techniques. For the reverse extrapolation process of RF threats, the threats were analyzed into a set of attributes, and the observed attributes were perceived at the receiver by using the modifying principle of the electromagnetic waves during transmission. The simulated threat data through uniform distributions were generated within the range of attribute values, and were compiled into a set of output to endow our agents with the reverse models of RF threats. Our agent's performance in the experiment proved that the agent's knowledge accumulated by compilation techniques was essential to threat identification and early warning, and to its continual survival in EW environments.

The final goal of this research is to repeatedly simulate various EW situations and for our agents to accurately identify the threat itself and its block diagram as well. In future work, we are implementing an integrated reverse extrapolation simulator, which consists of a module of communication between the transmitter of threats and the receiver of our agents for the generation of realistic attributes, a module of the block diagram presenting a certain operational principle of each threat, and a module of jamming techniques to test whether or not the identification of the threat is correct. We hope to be able to implement a fully autonomous agent to successfully identify RF threats as quickly as possible through our future work.

Acknowledgments

This work has been supported by the Electronic Warfare Research Center, Republic of Korea, under Grant EW41 "Reverse Extrapolation of RF Threats in Electronic Warfare Settings," 2013. We would like to thank our students, Jisu Ha and Cheolpyo Kim, for their help in implementing the machine learning algorithms.

References

[1] D.J. Bryant, F.M.J. Lichacz, J.G. Hollands and J.V. Baranski, Modeling situation awareness in an organizational context: Military command and control, in A cognitive approach to situation awareness: theory and application, eds. S. Banbury and S. Tremblay, Burlington, VT: Ashgate Publishing Company, Chapter 6. 2004.

[2] S. Noh and U. Jeong, “Intelligent Command and Control Agent in Electronic Warfare Settings”, International Journal of Intelligent Systems. Vol. 25, No. 6, 2010, pp. 514-528.

[3] J. Patrick and N. James, A Task-Oriented Perspective of Situation Awareness, in A cognitive approach to situation awareness: theory and application, eds. S. Banbury and S. Tremblay, Burlington, VT: Ashgate Publishing Company, Chapter 4, 2004.

[4] D.L. Adamy, EW 101: A First Course in Electronic Warfare, Artech House Publishers, Chapter 5. 2001, August 28, 2015

15:58 WSPC/ws-ijitdm ITDM20150828

[5] A. Golden Jr., Radar Electronic Warfare, AIAA Education Series, Chapter 2, 1988.

[6] B.R. Mahafza, Radar Systems Analysis and Design Using MATLAB, 3rd edition, CRC Press, Chapter 8, 2013

[7] M. Patzold, Mobile Fading Channels, John Wiley and Sons, Chapter 7, 2002.

[8] J.H. Van Vleck, "The absorption of microwaves by oxygen", Physical Review, Vol. 71, p. 413, 1947.

[9] J.H. Van Vleck, "The absorption of microwaves by uncondensed water vapor", Physical Review, Vol. 71, p. 425, 1947

[10] R. Hanson, J. Stutz and P. Cheeseman, Bayesian Classification Theory, Technical Report FIA-90-12-7-01, NASA Ames Research Center, AI Branch, 1991.

[11] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[12] S.P. Lloyd, "Least squares quantization in PCM", IEEE Transactions on Information Theory, Vol. 28, No. 2, 1982, pp. 129-137.

[13] Q. Yang and X. Wu, "10 Challenging Problems in Data Mining Research", International Journal of Information Technology and Decision Making, Vol. 5, No. 4, 2006, pp. 597-604.

[14] L. Hamilton, Six Novel Machine Learning Applications, Forbes (2014), http://www.forbes.com/sites/85broads/2014/01/06/six-novel-machine-learning-applications/.

[15] A.E. Spezio, "Electronic warfare systems", IEEE Transactions on Microwave Theory and Techniques, Vol. 50, No. 3, 2002, pp. 633-644.

[16] J. Heikell, Electronic warfare self-protection of battlefield helicopters: A holistic view, doctoral dissertation Helsinki University of Technology, 2005.

[17] P. Clark and T. Niblett, "The CN2 Induction Algorithm," Machine Learning Journal, Vol. 3, No. 4, 1989, pp. 261-283.

[18] I.H. Witten, E. Frank and M.A. Hall, Data Mining: Practical machine learning tools and techniques, 3rd edition. Morgan Kaufmann Publishers, 2011.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Challenges of Electronic Voting - A Survey

Abdelwahab AlSammak[1], Alaa AbdElRahman, Tarek ElShishtawy[2] and AbouBakr Elewa[3]

[1] Faculty of Engineering (Shoubra) - Benha University
[2] Faculty of Computer and Information - Benha University

[3] Ministry of Foreign Affairs
abdelwahab.asammak@feng.bu.edu.eg, alaaomarster@gmail.com ,t.shishtawy@ictp.edu.eg, elewabakr@hotmail.com
http://www.feng.bu.edu.eg/

Abstract

Electronic Voting (e-Voting) is the most important application in e-Government and e-Democracy. Thanks to the rapid growth in the use of computers and advances in cryptography, it is a serious push for e-Voting because many people already have access to the Internet. e-Voting can be the fastest, cheapest, and most effective way to administer the election, count the votes, and report the results. The main purpose of this paper is to highlight the major challenges facing e-Voting systems, introduce different ideas to face those challenge from different countries, and to explore the advantages and disadvantages of those ideas. Each of the challenges presented in this paper must be taken into account in crafting a legal framework for e-Voting to prevent harm before balloting is concluded.

Keywords: *Electronic Voting, e-Voting, e-Voting Requirements, e-Voting Challenges, Anonymity, Privacy.*

1. Introduction

1.1 Historical Background

The birth of democracy was in Athens in the sixth century B.C. where the first form of electoral laws was introduced [1]. Since that time, electoral systems have been designed and developed according to the characteristics of the countries in democratic governments around the world. Voting systems have evolved in response to the problems and the needs of political systems [2].
In many countries, interest in e-Voting is growing very rapidly. The number of e-Voting experiments taking place is also growing with different approaches and motivations of each country. By closely studying these experiences, it is possible to learn new and interesting lessons, lead to different schemes, and create a valid e-Voting system.

E-Voting machines were in use in the Netherlands for 20 years, with nearly the whole population vote using one of the DRE (Direct Recording Equipment/Electronic) voting systems available to vote. The introduction of this technology in the 1980s was not preceded by a public debate. In 2006, 90% of all votes in the Netherlands were expressed on the computer [3].

The birth of democracy was in Athens in the sixth century B.C. where the first form of electoral laws was introduced [1]. Since that time, electoral systems have been designed and developed according to the characteristics of the countries in democratic governments around the world. Voting systems have evolved in response to the problems and the needs of political systems [2].In many countries, interest in e-Voting is growing very rapidly. The number of e-Voting experiments taking place is also growing with different approaches and motivations of each country. By closely studying these experiences, it is possible to learn new and interesting lessons, lead to different schemes, and create a valid e-Voting system

E-Voting machines were in use in the Netherlands for 20 years, with nearly the whole population vote using one of the DRE (Direct Recording Equipment/Electronic) voting systems available to vote. The introduction of this technology in the 1980s was not preceded by a public debate. In 2006, 90% of all votes in the Netherlands were expressed on the computer [3].

The idea of e-Voting was introduced in Estonia I 2001. Their vision was to introduce Vote-over-Internet (VoI) in uncontrolled environments. Although at first they thought VoI could be used in the 2002 elections, they had to wait until 2005 to be a real option VoI in local elections. The first objective of VOI is to increase the participation maintaining voter interest in voting and increasing the interest of the younger generation. The other objective is

to stay in touch with modern Information and Communication Technology (ICT) and facilitate voting [4].

In 2002, the first e-Voting was conducted in Japan. Since then, ten local governments have conducted a total of twenty cases of e-Voting. In Japan, after “e-Japan Strategy", which aims to build an e-Government, was released in January 2001; many e orts of an e-Government and e-Democracy have been attempted. E-Voting can be seen in this trend [5]. In Korea, the participation rate is declining, a fact lead some to find a way to increase the participation rate. But an increase in the participation rate does not necessarily promote the quality of the representation itself. Due to the disproportionate representation in society, it can also over-represent the group that has been over-represented while an under-represented group becomes more under-represented. Therefore, improving the quantitative representation only make sense if the qualitative representation is made at the same time [6].

1.2 **Definitions**

Election An election is a process to obtain accurate data, representing a set of participants' responses to a question [7].

Voting, Voting means the fact to freely express choices between alternatives known to the public, e.g. candidates [8]. Voting is the most fundamental act of our democracy. Votes are mandatory for expressing people's will, which must be both secret and restricted to only one per citizen. It should be secure enough, easy to register, easy to vote, and easy to count the votes. Voting systems should comply with the principles of non-discrimination and democratic elections [9].

Vote, A vote is that physically represents the response of a participant in a particular issue. A vote is a selection, usually from a predetermined set of responses called candidates. Sometimes a vote includes a selection, which is not a member of the predetermined list, and is called writing -in stations [7]. The vote is the most powerful tool to express the content and citizen control over government agencies. The vote should not be understood as a mechanical process, but as having a capacity to create its own, because it provides unification of the people. Although the act of voting is considered a personal right, the process engages the development of the nation as a whole. The choices of procedures and tools in place to support the “unification" are of vital importance because they must respect the creative capacity of the unification process, without introducing disparities [9].

Ballot, One or more votes are grouped in a structure called a ballot. Each question in an election is called a race, so each race has a set of candidates potentially receive the votes of electors [7].

E-Voting, e-Voting is a term encompassing several types of voting, includes both electronic means of casting a vote and electronic means of counting votes. E-Voting technology can include punched cards, optical scan voting systems, and specialized voting kiosks. It can also involve transmission of ballots and votes via telephone, private computer network, or the Internet [10]. E-Voting types fall into two major categories:

- On-site e-Voting (supervised by representatives of governmental with e-Voting machines at the polling station)

- Remote e-Voting (not physically supervised like voting using computer via the internet, using mobile phones via SMS, or at public kiosks).

Electronic elections are conducted either using DRE machines or over the Internet. Although DREs have benefits such as speedy results, accuracy, reduction in manpower and paperwork, they are vulnerable to sabotage and equipment malfunction. Further, if a malfunction is detected, there seems no way to conduct a recount and the only remedy is a recast of ballots. Internet voting provides ease of access and eliminates absentee ballots, but is surrounded by many more security concerns than the DRE systems. Figure 1 illustrates the flowchart of classical voting, on-site e-Voting, and remote e-Voting, in the same time fig 2, illustrate how I-voter can vote using any device connected to the internet.

Technology is very promising to serve as a mean to cope with the crisis of participation and confidence that democracy faces todays [11, 12]. For example, it can be used to make democracy more accessible to citizen as e-Voting can provide great opportunities for improvement of certain groups access to the electoral process. The following groups are eligible [2]:

- Visually impaired citizens can use a headset connected to DRE or the PC if Internet voting is used. - Minorities can access e-Voting systems in their preferred language.

- Citizens living and working abroad can vote online from their own homes. Citizens who cannot attend at a polling station to cast their votes can vote Online from their own homes.

2. Requirements of e-Voting

Requirements of traditional voting (paper ballot) are also valid for e-Voting like complying with the principles of non-discrimination and democratic elections. Any voting system must meet the following requirements [13, 14, 15].

Universality, all voters have the right and ability to vote using the system

Authenticity, only eligible voters can participate.

Uniqueness, No voter should be able to vote more than once.

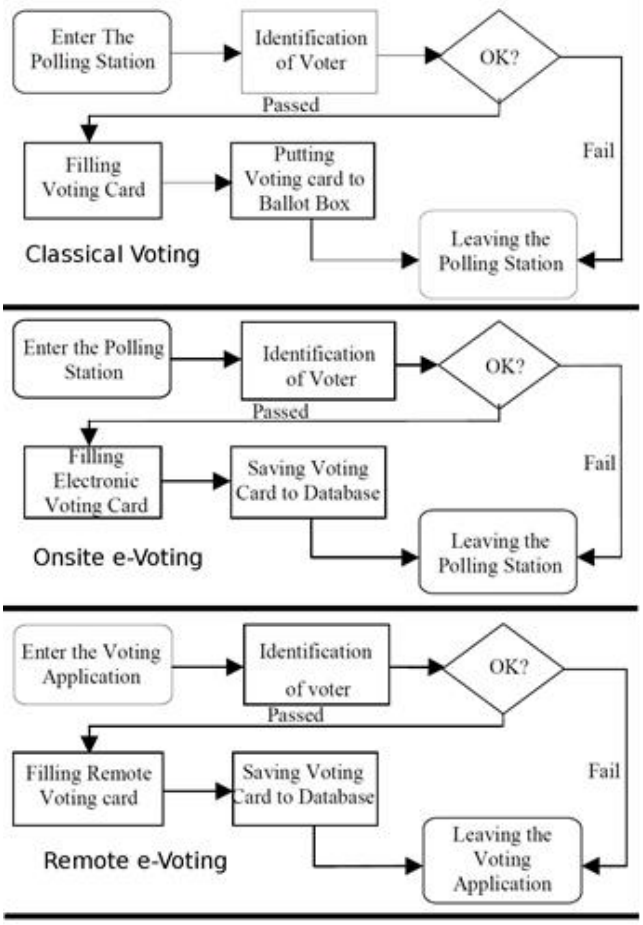


Fig. 1. Classic voting vs. On-site e-Voting vs. Remote e-Voting

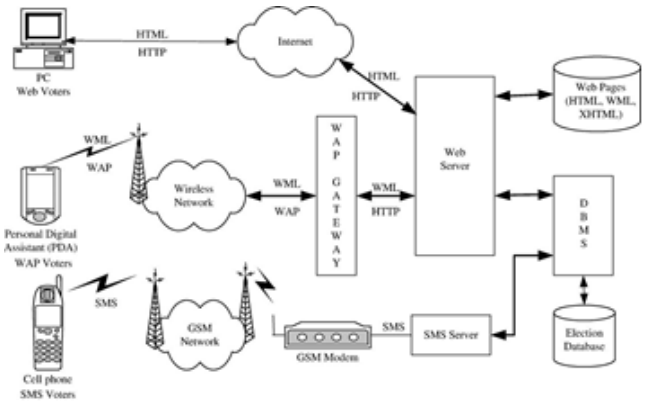


Fig. 2. An example remote e-Voting system

Reliability, the system should function without compromising votes, even if system failure occurs

Accuracy, the votes are properly recorded.

Integrity, Votes cannot be edited or deleted.

Flexibility, the system should be usable by different types of voters (support multilingual voting ballots, accommodate disabilities by audio or visual features, support different input methods, etc.).

Convenience, Electoral systems should not require additional skills to be usable without unreasonable need for equipment.

Transparency, Voters should be able to understand the overall system.

Secrecy, Votes should be secret and a voter must not have a record of voting choices.

Anonymity, Each voter has the right to cast his vote secretly, and no one should be able to relate a voter to his/her vote.

Freedom/Uncoercibility, The citizen must be able to vote without being forced by the government to vote for a particular candidate.

Audit/Accountability, The system has the ability to verify that votes are properly counted.

Verifiability, the system must be tested by election officials.

Cost, the system should not be too expensive.

A voting protocol is said to respect privacy when an intruder cannot detect if arbitrary honest voters VA and VB swap their votes. In general, this means that the intruder cannot know anything about how VA (or VB) voted. This can be expressed as follows [16, 17]:

$$S\,[VA\ \{a/v\} |\ VB\ \{b/v\}] \approx S\,[VA\ \{b/v\} |\ VB\ \{a/v\}]$$

Even if the result of the election is necessarily revealed, the definition above is still robust [18].

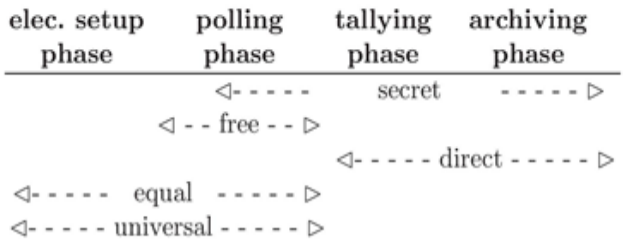


Fig. 3. Election Phases and requirements

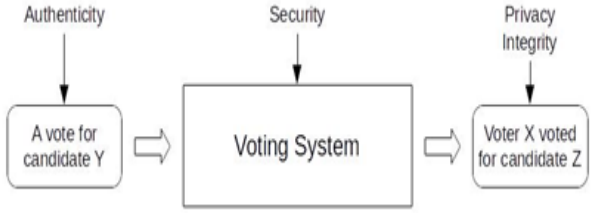


Fig. 4. Requirements of Voting

Unfortunately, some satisfying some requirements contradicts with satisfying others. This results in some challenges. In the following section, challenges of e-Voting systems are introduced. Voting in Egypt is like any other country; most countries still using the traditional voting technique to elect the government, but the Egyptian government now thinking about electronic rather than conventional Vote, to avoid the problems they are facing. Where there are many problems in the conventional voting system used in Egypt, like:

1. Relation between the government and the people usually suffers from lack or trust.

2. Sometimes, government coerces and carries on the voters to vote for a particular candidate, and eliminate them from voting freely.

3. Some candidates trying to win by buy the votes from the voters.

4. Government can cheat by substitute the original ballot by derivative ones.

Therefore, there must be another way to solve these problems or reduce it as much as possible, and give the voters the confidence to believe in the system. Consequently, new technologies must be used to improve the election process by building new systems that are more convenient to people [13].

3. Challenges of e-Voting

It is important to control and to observe different stages of the election process. It is necessary to be able to guarantee the well-functioning of the system before the start of the election period, during the voting period and afterwards. This means that there must be a focus on certification processes before the processing of the data actually begins as well as on proper mechanisms for post-auditing of the elections [19, 20].

3.1 Legal Challenges the fact that municipalities are legally obliged to keep a registry of eligible voters is certainly also favorable for any e-Voting system [21]. Law and consequently the constitutional law de ne clear and strict regulations for voting and instruments used. To use the computer-aided communication in these fields, used techniques must satisfy the relevant legal requirements [22]. Any attempt to introduce e-Voting, i.e. a voting process, which enables voters to cast a secure and secret ballot over the Internet or an Intranet, will have to address a series of complex constitutional and legal issues. Our paper refers to these democracy-oriented legal and constitutional requirements, which every electronic voting system has to comply with [11]. On the other hand, law can provide legal protection to e-Voting systems. Attacks against mission-critical systems in countries like the United States and the United Kingdom are treated as criminal cases for which the perpetrators must be prosecuted. The act of hackers/crackers unauthorized access to a computer system can be compared to someone breaking into a home as a way to check if it is secure [23, 24].

Voter Identification; Identification and authentication of the voter when e-Voting is used at a polling station, the voter identification process may remain the same, but it can also change if an electronic register of voters used. In this case, arrangements should be put in place to ensure that the identity of the voter cannot be linked to his/her voice. If biometric features have been used for the registration process, these features can be used for voter authentication. Vote home Internet is different and an electronic remote identification system must be developed. Voters could authenticate with an electronic identity card having voter credentials or, if such a system exists, authenticate using a combination of user name and password with a control issue (e.g., date of birth). It is important to realize that without a physical token, authentication of the voter is less reliable and it is much easier to sell his vote in disclosing the user name and password to a third party. It should be noted that when voters must make their own user name and/or password (for example, when registering to vote), they may forget or misplace the username and/or password. Thus, a system must be established to provide a username and/or password in the short term while at the same time as the voter can vote only once [2, 25, 26].

Verifiability; is a central institution of modern e-Voting systems. Intuitively, verifiability means that voters can verify that their votes were counted and the election published result is correct, even if the voting machines/authorities (partially) unreliable [27].

Maintaining Anonymity to preserve the secrecy of the vote as one of the main principles of democratic elections, it is important that at some point in the voting process, the link between the identity of the voter and the vote itself is divided (which is also known as unlink ability). This should preferably be done immediately after the voter has cast his/her vote. Since the vote and the voter should not be linked, it is important to establish an administrative procedure that has access to register to vote and voter lists (preferably managed by different authorities), when and under what circumstances they will access, how long records exist, and how and by whom they will be deleted. In case of reversible vote, specific technical solutions must be implemented [29, 30,31].

3.2 Social & Cultural Challenges

E-Voting was introduced in Belgium in 1994. Ironically, no action had been taken to determine the opinion of facing this original method of voting constituents. In [32], the social and empirical dimensions of the legitimacy of this new method through several empirical indicators used in an investigation on the occasion of May18, 2003 federal election: (A) It was difficult for voters to vote for a computer; (B) the extent to which they trust to vote on a computer; (C) if they have a philosophical/social opposition to a vote on a computer [32].

The result of this provision inserted in the Belgian electoral law by the Act of 11 April 1994 was the introduction of a voting system of the computer in a growing number of municipalities for all elections in Belgium since the 19942 system used in Belgium is separate from Internet voting and voting by computer network. Voters go to the polling station where they are asked to vote on the computer. The objective of the system is to make the voting and vote counting easier and faster [32].

Paradoxically, this new method of voting had not yet been evaluated in depth. In particular, no action had been taken to determine the opinion of facing this original method of voting constituents. For this reason, during federal elections of 18 May 2003, Belgium, the authors conducted a large survey of voters leaving the polls to determine the views of the Belgians on e-Voting immediately after using this new technique of vote. Two main issues were considered: (a) the extent to which e-Voting, as used in Belgium is considered easy or di cult to use, and

(b) If e-Voting is socially accepted or rejected by voters who use? For example, it should be noted that 20.37% of voters without education considers that e-Voting is `di cult' or `very di cult'. There is a digital divide to consider, even if it is not striking [32].

In December 2006, the federal and regional governments have asked a consortium of seven Belgian universities to present a study on the legal technical, Organizational, socio-political and a range of voting systems. In addition, special attention should be given to the accessibility and usability of the system for people with disabilities [33].

A research team of the Universit libre de Bruxelles (ULB) verified the legitimacy of e-Voting in 2003; the main conclusion was that 88% has a favorable attitude towards the system, while 8.5% unfavorable (the rest 3.5% had no opinion or did not answer).

More detailed results are as follows: 95% of voters easily find the system very easy to use and 85% have no problem in principle with e-Voting, and 89% are fully con dent or somewhat con dent in the e-Voting system [33].

A majority of voters who had confidence in e-Voting also expressed his confidence in the ballot, but more moderately. On the other hand, those who dis-trusted the new method are those who have contributed most ballots. While the Belgian e-Voting system is not as vulnerable as DRE used in countries such as the United States, the Netherlands, and France. Indeed, this computer can overwrite a vote and subsequent verification of magnetic stripe cards will not reveal such an attack. In addition, the central production and distribution disk requires a complex chain of custody in which the focus should be trusted (eg, how can someone be sure that the software is released later

the same as the software running on each computer station?) [33].

Furthermore, some irregularities have been reported, for example, with respect to the management of passwords to activate devices. Voting machines are themselves \stupid" machines that do not store the ballots, but it is not in-conceivable that the material of such a machine can be changed to attack the privacy and/or the integrity of the vote. These machines can also be vulnerable to side channel attacks, for example, based on the electromagnetic analysis. As these machines are quite large and not useful for any other purpose, it is unlikely that they are stored in a place high physical security. The paper-based system is not without flaws, namely: the authors were informed (off the record) that frauds are known in which votes for specific candidates are added during the counting process. It is also very easy to spoil a ballot by an additional mark on it [33].
Voters were able to vote by electronic ballot box instead of throwing a bulletin plain paper in a traditional urn. The electronic ballot box was designed as a computer with a touch screen, like a terminal Mini Bank. Ballots submitted electronically were counted as regular ballots in the election. Experience Oppdal was more comprehensive than the rest. Oppdal municipality has nearly 5,000 voters, and electronic option was available in all seven areas of the town vote [4]. 91% of the electorate voted electronically to the election Svalbard. In the municipality of Bykle 53% opted for the method of e-Voting. In Oppdal 34% of the electorate voted electronically. In the district a voting Larvik, stre Halsen, 18% opted for electronically [4].

Family Voting Concerning the problem of “family voting" and similar possible in influences on the individual voters decision, which represent a major criticism of the use of internet voting, it was brought up that postal voting suffers theoretically from the same problem and that there exist means to guarantee the voters expression of free will (e.g. by introducing the possibility to recast the vote when it was cast via internet).
Vote Buying Any person who purchases or offers to purchase a vote of any elector at an election by the payment of money or the promise to pay the same at any future time, or by donation intoxicating liquor or anything else of value, are considered guilty of an offence. Every voter in an election that takes or receives money or other thing of value, provided that the same shall be paid at any time in the future in exchange for voting as an elector for a particular candidate, or promise to vote for a particular candidate, is guilty of an offence [34]. Coercion and vote buying: These risks are significant as it is impossible to prevent situations in which the voter casts a vote under pressure, or proves to a third party whom she/he has voted for [33]. Receipt-freeness are necessary to prevents vote selling/buying, ensuring that voters are not used as a proxy to cast votes [8, 35.36].

3.3 Technical Challenges

Design Flaws; An important decision when defining a strategy for e-Voting is whether to use open-source or proprietary software. This is particularly relevant to the question of trust. Many companies use proprietary e-Voting software, which has the disadvantage that in most cases, the rights holder is not the source code available to the general public (or makes available partially or temporarily) [2, 37].
In [35], security analysis of the source code of Diebold AccuVote-TS 4.3.1 has been introduced. It is one of the first electronic machine paperless voting systems used in a large market share. It is based on Windows CE and is developed in C++. The analysis shows that this voting system is far below even the most minimal security standards applicable in other contexts. Several problems including unauthorized privilege escalation, incorrect use of cryptography, vulnerabilities to network threats, and poor development process software were identified. Without any insider privileges, can make unlimited votes without being detected by mechanisms in the terminal software to vote. In addition, Even the most serious of our outsider attacks could have been discovered and executed without access to the source code. Faced with these attacks, the usual worries about insider threats are not the only concerns; foreigners can do damage. The insider threat is also quite considerable, showing that not only can an insider, such as a poll worker, modify the votes, but that insiders can also violate the privacy of voters and results of the votes with the voters who cast. This voting system is unsuitable for use in a general election. All e-Voting system paperless could suffer the same flaws, despite any “certification" it would have been other-wise. In [35], it was suggested that the best solutions are systems with an “audit trail voter verifiable," where an e-Voting system might print a ballot that can be read and verified by the voter votes.

Spoofing; Sites spoof malicious Web sites that are created to look like legitimate Web site, in a scenario of voting it is understood that this could be really bad, the site could be used to launch phishing attacks to collect credentials voters as a PIN or password required to vote. The website may look exactly like a voting site in the state, but redirect the browser to the voter to a malicious Web server. There are many ways that an attacker could spoof a legitimate site vote. One way might be to send emails to users tell users to click on a link, which then set up a voting site were false adversary could collect the credentials of the

user, to steal the vote, and then use it to vote differently. An attacker could also establish a connection to the legitimate server, feed the user a fake web page, acting as a man in the middle, transfer, and control all tracks between the user and the web server. Transferring information between the user and the server, the user's voice can be changed before further sent to the server. [7, 10].

Malicious Payload; Threats of a modern system of e-Voting security: Malicious payload is a threat to the security of the personal computer of the voter. The malicious payload is software or configuration to damage and could be a virus, worm, Trojan horse or a remote control program that is perhaps the greatest threat in a scenario of voting. If a malicious program is installed on the computer of the voter, it could change the secret ballot. The owner of the computer may not be aware of even have one installed because these programs can be di cult to detect (run in stealth mode) malware. Malware of this kind have increased in sophistication and automation in recent years in a way that they can do more damage, more likely to succeed and to dress better. Even if a system of Internet voting has strict protocols for encryption and authentication, malicious code can do its damage before the other security features are applied to the data. [10].

4. Attacks

E-Voting technology can speed the counting of ballots and provide better accessibility for disabled voters. However, e-Voting could also facilitate electoral fraud [10]. Internet based voting systems require strong safeguards against hacking attacks, viruses and Trojans. Software continues to get complex and can never be bug free. A virus or network attack can also be mounted during the verification process and result in false positive verifications. Network attacks may be met by cryptographic key exchange and distributed back-end databases. Information dispersal algorithms and verifiable secret sharing schemes may be used to maintain system fairness such that no single server stores all the cast ballots and the partitions are distributed over independent servers. As long as a majority of these servers remains honest, the possibility of sabotage remains low.
Initiated in a voting system may include hardware vendor and/or pre made software, election officials, poll-workers, maintenance technicians, and others. It is impossible to completely prevent internal attacks, but levels of resistance to such attacks systems [36].
Attack on e-Voting system can be classified according to the configuration; attacks such as advertising, pro t attacks, terrorist attacks and attacks that are motivated by the desire to create instability in the state government and more. Because of the safety pro le at high risk of e-Voting, it is necessary that each component or unit in the electoral process presents the principles of security (confidentiality, integrity, availability) and controls must be applied to protect them. E-Voting requires the implementation of protective measures to fight against all identified threats and the ability to prevent unregistered. An attacker must have three things:
Reason The reason for wanting to attack.
Possibility time and access to a full attack.
Method the knowledge and tools necessary to perform an attack skills.

Attack on an e-Voting system can be classified according to the model; these attacks are attack ads, non-profit force attacks and terrorist attacks are motivated by the creation of the instability of the current government/democracy. Threats could be, for example internal vendor, election officials. Alternatively, they can be external, such as individuals, organizations and funded, states, parties, criminals, terrorists, many of whom cannot even be prosecuted. The motivations of attackers ranging from advertising, foreign intelligence and terrorist acts, governments handling system to their advantage [9].

Voting systems based on the Internet are vulnerable to attack by three main points; the server, the client and the communication infrastructure. Penetration attacks target the client or server directly while DoS attacks target service and interrupt the communication link between the two. The penetration attacks involve the use of a distribution mechanism for carrying a malicious payload to the target host in the form of a Trojan horse program or a remote control. Once executed, it can spy on the ballots, prevent voters from casting ballots, or worse, change the ballot according to his instructions. Remote control software can compromise the secrecy and integrity of the ballot by those who monitor the activity of the host.
In the context of new voting technologies (NVT), piracy is seen as an entry in the illegitimate system made by anyone external to the process management. For DRE voting systems and scan ballots, safeguards must be put in place to prevent physical handling with appliances. Election Observation Mission (EOM) must check, for example, the USB ports or other external connections are not easily accessible. In addition, the storage and transport of NVT devices must be conducted in the context of secure protocols defined manner, and access to peripherals must be observed when they are not in use, with appropriate records kept. Hacking can also occur if the devices are connected to the Internet [37].
Another challenge is the need to preserve the secrecy of the vote, while at the same time the integrity of the results. It has hitherto been di cult for e-Voting process - especially Internet voting - to meet these two fundamental principles of democracy at the same time. Another

challenge is that NVT present additional difficulties in the electoral process, such as the need to amend the legislation; planning how NVT will be acquired, tested, evaluated, certified and secure; and provide education and training of electoral agents voter; and general as to the transparency of the process and access concerns for observers. Using NVT therefore not necessarily built trust; rather, it seems to require existing confidence in the administration of elections for successful implementation. These challenges, if they are not fully taken into account, can weaken public confidence in the electoral process.

In addition to physical intrusion, external hacking is a particular threat. The EOM should check how the system prevents or detects an illegitimate access, and should assess the likely effectiveness of these measures. In Internet voting systems, the EOM should consider how the system verifies the identity of the voter and the threats that could create potential. In addition, the overall protection of information from unauthorized access, through the use of trans-mission lines dedicated firewall and overall concepts of external security access systems, should be considered. Data manipulation by officials, suppliers or electoral technicians is another potential threat posed by NVT. The EOM should ensure that procedures are in place to limit the ability of any person to undermine the system. For example, there should be a division of labor within the electoral administration to minimize the possibility of an internal manipulation. The physical and electronic access to the NVT system must be strictly regulated by written procedures. Any access should be limited and observable so that election officials or suppliers have access only to components that fall within the scope of their responsibilities. The EOM should also check if sensitive system operations are performed by more than one person and a record of all transactions is maintained. Safety procedures must be both effective and fully implemented; the measures that bring evidence justified as inviolable security seals with unique numbering, secure stamp documents and similar mechanisms to prove the authenticity of procedures to provide security against malpractices. Although these security measures are necessary, they may not be sufficient to ensure electoral integrity or to maintain public confidence. Appropriate verification measures, including audits of voter -verified paper documents are needed to fully guarantee the integrity of the vote [38].
In practice, the two most important problems of computer security compromise and coercion. Cryptography cannot protect a voter coercion when voting from home or a public place, but the system must include features to prevent coercion. A solution to the problem of stress is the ability to submit multiple ballots. The system allows the voter to multiple re-vote the final ballot. It is also possible to vote at the polling station and a ballot crush any e-Voting, no matter what the time stamp (submitted before or after e-Voting shown). The previous question is a measure against coercion external attackers, but stress can also be done by election insiders/employees. A voter authenticates before launching a ballot, and election officials with access to the authentication sys-tem can detect any electronic e-Voting by a voter. The election officer cannot see the contents of the ballot cast, but he could see/detect constraint voter casting a new vote. An election official coercing access to ballots counted could also verify that the voter constraint (s) (the victim (s)) does not have to vote again.
If forcing the voter (s) to submit a ballot with the desired effect coercion can observe the counted ballots and verify if the ballot (s) are present among them. The compromised computers come home is another major threat. A significant fraction of home computers is compromised, and the Norwegian protocol must provide the voter an opportunity detect falsification ballot without relying on computers. It is complicated, because the voter cannot perform cryptographic calculations without a computer, and this was the method of using a generator and receiver codes pre-generated receipt is involved. The voter receives reception codes pre-calculated on his voting card with his voting card and after casting a ballot, receiving codes generated by the generator and receiving the ballot box is sent to the voter by not the system and the computer, but through an independent channel (postal service). If the computer voter is corrupt, the attacker may be able to see the ballot, and the attacker can also change the ballot. There-fore, these security mechanisms allows the voter to note manipulation with high probability [10].

4.1 Physical Attacks

Many physical attacks can be made on the e-Voting system to sabotage the election. Vandalism of e-Voting systems makes it unusable for Election Day. Saboteur can remove network connections and pull the plug on e-Voting systems causing lost votes. Attackers can remove hard disks or smart cards to replace falsified data. E-Voting machines could be stolen by attackers discover information confidential voting on users [23].

4.2 Overloading Attacks

Denial-of-Service (DoS) Attack Distributed Denial-of-Service (DDoS) attack is an attack on a computer system or a network in which a simple auto-mated request is repeated at a very high frequency, with the aim of overloading the connecting lines of the system or the calculation of capabilities. These attacks are detectable and may require the postponement of the election. EOM should therefore check what security measures were put in place to protect systems against such attacks [39].

DOS attacks are performed by automatically sending a flood of messages on a website, server, or on a channel similar to crash or reduce the quality because it cannot handle all the traffic generated. Using a DOS attack distributed (DDOS), attackers can cause routers to crash or electoral servers being flooded, or it is possible to attack a large number of hosts such demographically targeted to stop the operation of the election. This can be a major threat to Internet voting if such voting takes place in one day. It is important to have additional bandwidth to handle the traffic and some voting systems I will describe later, the vote may occur over several days in advance of the election [10, 40, 41, and 42].

Ping of Death The ping of death relies on a flaw in some Transmission Control Protocol, Internet Protocol (TCP/IP) stack implementations. The attack relates to the handling of unusually and illegally large ping packets. Remote systems receiving such packets can crash as the memory allocated for storing packets over flows. The attack does not affect all Systems in the same way, some systems will crash, and others will remain unaffected [23].

Packet Flooding Packet flooding exploits the fact that establishing a connection with the TCP protocol involves a three phase's handshake between the systems. In a packet flooding attack, an attacking host sends many packets and does not respond with an acknowledgement to the receiving host. As the receiving host is waiting for more and more acknowledgements, the buffer queue will fill up. Ultimately, the receiving machine can no longer accept legitimate connections [23].

4.3 Receipt Attacks

Trash Attack The idea of the trash attack is that if voters throw away their (paper) receipt, then authorities who find these receipts could conclude that these voters will not check their receipts on the bulletin board, and hence, ballots of such voters can safely be modified [27].

Clash Attack; the simple idea behind the shock attack, is as follows. Voting machines are trying to provide different voters with the same reception, where the name of the attack. Accordingly, the authorities can safely replace the ballots news on the scoreboard; therefore, manipulate the election without being detected. In [27], it was shown that, surprisingly, many e-Voting systems that have been designed to provide the verifiability between systems that have been used in real elections are vulnerable to this attack, under realistic assumptions of trust in machines and authorities vote. Our results show that this attack is a potentially dangerous attack for a large class of e-Voting systems. It must be noted that the shock attack can work even if the voters and election observers know exactly how and what the electorate voted. So confront attacks are different and more subtle than the known ballot stuffing attacks (see, for example, attacks ballot stuffing) [27].
This attack does not seem to have attracted much attention in the literature. Even if the attack is quite simple, under reasonable assumptions confidence, it applies to several e-Voting systems that have been designed to provide verifiability. In particular, it applies to large as well as two e-Voting systems that have been deployed in real elections and voting systems Three Ballot and Vote/Anti-Vote/Vote (VAV), the Wombat voting system and, its alternative voting system Helios [27].

4.4 Man-in-the-Middle Attack

Fraud in the form of fake servers must also be taken into account. Some server may pretend to be the official server by tampering with the DNS or by using a name very similar to that of the official server (Man-in-the-Middle). To protect the system against Man-in-the-Middle attacks, a digital signature may be applied to the ballot to ensure verification of the voter submitting the ballot. However, it is of utmost importance that the confidentiality of the vote is not threatened [4.43].

5. Conclusion and Future Work

From the previous section, it seems that online voting is very promising for application in Egypt despite of the existing challenges. Implementing an online voting system offers many advantages. One of the most important advantages is its ability to increase voter turnout by making the elections more convenient and more accessible to busy voters, lazy voters, and voters with special needs. Other advantages include the low cost, ease of administration, and auditability. The difficulty of applying online voting in Egypt lies in convincing voters that their privacy is maintained at all times. Public must be informed about the manner by which the Internet is protected from outside influences, including national and international hackers as well as whom might try to cast more than one ballot. In addition, online voting may require some legal regulations to be applied in Egypt.
We are working on developing a complete end-to-end auditable online voting system that is capable of satisfying all the requirements of e-Voting, getting over the technical challenges, surviving against possible attacks, complying with legal regulations, and gaining the confidence of the Egyptian people.

Acknowledgement

I would like to express my appreciation for the support I got from Egyptian Ministry of Foreign Affairs and all the help I received from George Town University Library;

they gave me the chance to access the latest publications in the field.

References

[1] J. Pujol-Ahull, R. Jard-Ced, and J. Castell-Roca, "Verification systems for electronic voting: A survey," pp. 163–177, 2010.
[2] S. Caarls, E-voting Handbook: Key Steps in the Implementation of E-enabled Elections. Council of Europe, 2010.
[3] A.-M. Oostveen, "Outsourcing democracy: Losing control of e-voting in the Netherlands," 2010.
[4] "Electronic voting - challenges and opportunities," 2004.
[5] M. Iwasaki, "E-voting in japan: A developing case?" in 4th International Conference of Electronic Voting, vol. 167, 2010, pp. 283–295.
[6] H.-W. Lee, "Political implications of e-voting in korea," vol. IX, no. 1, pp. 91–107, 2005.
[7] M. F.M.Mursi, G. M. R. Assassa, A. Abdelhafez, and K. M. Abo Samra, "On the development of electronic voting: A survey," vol. 61, no. 16, pp. 1–11, 2013.
[8] L. Fouard, M. Duclos, and P. Lafourcade, Survey on Electronic Voting Schemes, 2007.
[9] D.Zissis, "Methodologies and technologies for designing secure electronic voting information systems," 2011.
[10]M. Stenbro, "A survey of modern electronic voting technologies," 2010.
[11]L.Mitrou, D. Gritzalis, S. Katsikas, and G. Quirchmayr, Chapter 4 E-VOTING: CONSTITUTIONAL AND LEGAL REQUIREMENTS AND THEIR TECHNICAL IMPLICATIONS, 2009.
[12] Eric A. Fischer and Kevin J. Coleman, "The direct recording electronic voting machine (DRE) controversy: FAQs and misperceptions," 2005.
[13] M. Abo-Rizka and H. Ghounaim, "A novel in e-voting in egypt," vol. 7, no. 11, pp. 226–234, 2007.
[14]Abdalla Al-Ameen and Samani A. Talab, "E-voting systems security issues," vol. 3, no. 1, pp. 25–34, 2013.
[15]T.R¶ssler,"E-voting: A survey and introduction," 2009.
[16] R. Kofler, R. Krimmer, and A. Prosser, "Electronic voting: algorithmic and implementation issues," in Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003, 2003, pp. 7 pp.–.
[17] M. Bishop, "An overview of electronic voting and security," 2003.
[18] S. Delaune, S. Kremer, and M. Ryan, "Verifying properties of electronic voting protocols," in Proceedings of the IAVoSS Workshop On Trustworthy Elections, 2006, pp. 45–52Proceedings of the IAVoSS Workshop on Trustworthy Elections. 45-52, 2006
[19] Jan Gerlach and Urs Gasser, "Three case studies from Switzerland: E-voting," 2009.
[20] Bryan Schwartz. (2013) Establishing a legal framework for e-voting in Canada.
[21] I. Ray, I. Ray, and N. Narasimhamurthi, "An anonymous electronic voting protocol for voting over the internet," in Advanced Issues of E-Commerce and Web-Based Information Systems, WECWIS 2001, Third International Workshop on., 2001, pp. 188–190.
[22] P. Heindl, "E-voting in Austria legal requirements and first steps." pp. 165–170, 2004.
[23] A. Al-ameen and S. Talab, The Technical Feasibility and Security of E-Voting, 2011.
[24] Lilian Mitrou, "ELECTRONIC VOTING OBSERVATORY II VOTOBIT," 2004.
[25] Christine Lai, "The impact of voter identification laws on voter participation," 2013.
[26] Rodney Smith, "Multiple voting and voter identification," 2006.
[27] R. Kusters, T. Truderung, and A. Vogt, "Clash attacks on the verifiability of e-voting systems," in 2012 IEEE Symposium on Security and Privacy (SP), 2012, pp. 395–409.
[28] O. etinkaya and D. etinkaya, Anonymity in E-Voting Protocols, 2008.
[29] Robert Krimmer and Melanie Volkamer, "Observing threats to voter's anonymity: Election observation of electronic voting," 2006.
[30] D. Chaum, P. Y. A. Ryan, and S. Schneider, "A practical voter-verifiable election scheme," in Computer Security ESORICS 2005, ser. Lecture Notes in Computer Science, S. d. C. d. Vimercati, P. Syverson, and D. Gollmann, Eds. Springer Berlin Heidelberg, 2005, no. 3679, pp. 118–139.
[31] L. Langer, A. Schmidt, M. Volkamer, J. Buchmann, and T. U. Darmstadt, in Classifying Privacy and Verifiability Requirements for Electronic Voting, 2009, pp. 1837–1846.
[32] P. Delwit, E. Kulahci, and J.-B. Pilet, "Electronic voting in belgium: A legitimised choice?" vol. 25, no. 3, pp. 153–164, 2005.
[33] D. D. Cock and B. Preneel, "Electronic voting in Belgium: Past and future," in E-Voting and Identity, ser. Lecture Notes in Computer Science, A. Alkassar and M. Volkamer, Eds. Springer Berlin Heidelberg, 2007, no. 4896, pp. 76–87.

[34] Michael Ian Shamos, “Electronic voting glossary,” 2011.
[35] T. Kohno, A. Stubblefield, A. Rubin, and D. Wallach, “Analysis of an electronic voting system,” in 2004 IEEE Symposium on Security and Privacy, 2004. Proceedings, 2004, pp. 27–40.
[36] J. Epstein, “Internet voting, security, and privacy,” vol. 19, no. 4, p. 885, 2011.
[37] R. Krimmer, “Handbook for the observation of new voting technologies,” 2013.
[38] M. Volkamer, Evaluation of Electronic Voting - Requirements and Evaluation Procedures to Support Responsible, 2009.
[39] Sandeep Mudana, “Security flaws in internet voting system,” 2004.
[40] Darshan Lal Meena, “Effects of DoS attacks on the e- voting system and feasible measures to prevent them,” vol. 3, no. 4, pp. 16–21, 2014.
[41] Volkamer, M.: Evaluation of Electronic Voting - Requirements and Evaluation Procedures to Support Responsible, 2009
[42] Electronic Pool Book systems as distributed Systems: requirements and Challenges (national Conference on state certification testing of Voting systems may 19-20, 2015, WA, US.
[43] Issues and challenges of transition to e-voting technology in Nigeria, International Knowledge Sharing Platform, Public policy and administration research 2015

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Design of a Portable Random Access Wireless Network Transmitter

Rashid Hassani[1], Prabhu Gudapusetty[2] and Peter Luksch[3]

[1] Department of computer science, University of Rostock
Rostock, Germany
rashid.hassani@uni-rostock.de

[2] Department of computer science, University of Rostock
Rostock, Germany
prabhu.gudapusetty@uni-rostock.de

[3] Department of computer science, University of Rostock
Rostock, Germany
Peter.luksch@uni-rostock.de

Abstract

There is intensive attention on improving random media access control protocols in wireless environments. This paper proposes low-complexity, portable and flexible development of the wireless transmitter employing Random Access (RA) protocol. To develop this RA solution, the software architecture is roughly divided into three modules. The host software (i.e., micro-PC interface), Universal Serial Bus (USB) and Radio Frequency (RF) module. In our transmitter based scheme, the generated packets move from the higher layer (i.e., MAC layer) to the physical layer and then transmit over the air. At the other end, the packets will be received by the evaluation board, which is acting as a receiver. By using packet sniffer tool, we are able to sniff the radio packets from the micro-PC based RF-fronted wireless transmitter. Our results have been conducted through various test scenarios and emphasize the validity of our development in which, by our new developed RA solution, the generated radio packets are transmitted over the air successfully without any packet loss.

Keywords: *Random access, USB, Wireless network, Wireless transmitter.*

1. Introduction

The design of random media access control (RA MAC) protocol is widely considered as critical issue in wireless environments and currently is considered as an active research area for challenged environments such as the satellite or wireless sensor networks. Since, the number of devices capable of interconnecting is steadily increasing, if the contention among different users in shared medium is not appropriately controlled, this may lead to large number of collision, resulting in wastage of resources such as bandwidth and energy as well as system efficiency [1][2]. This leads to a rise of many interesting research questions on how to manage the shared medium efficiently. To improve the overall quality of service at the user end, there are various random media access control protocols which can be used. For example, for a large set of users, a distributed wireless MAC protocol is preferred, (i.e., a Random Access (RA) protocol). In RA protocols (e.g., ALOHA), the medium is accessed without coordination between the users, leading to possible collisions. RA protocol schemes are suitable for handling initial access, burst traffic and short packets in up-link satellite communication. However, the collision, propagation delay and packet loss are some series of pitfalls of this technique which may vary between different transmitter-receiver pairs [3][10][19]. This leads to a rise of many interesting research questions on how to manage the shared medium efficiently to avoid or resolve collisions and of course packet loss. Multi-User Detection (MUD) is a receiver based scheme where multiple transmitters are transmitting at the receiver end. The user data is separated based on a signature waveform [4][18]. However, in RA, the active number of transmitters at any particular time may not be known at the receiver. It is necessary to first know the active number of users involved in current transmission from the received signal, followed by MUD based on the signature of the waveform. Multi-packet reception reduces the collision because collided packets can be recovered by separating them from overlapping packets by using signal processing techniques [5][6][20].

The contribution of this work is to develop the RA solution which is a portable transmitter with wireless capabilities with proper connection to micro-PC and a radio frontend working in the WiFi frequency bands with ability to support various RA schemes (e.g., ALOHA, slotted ALOHA, etc.). This can be achieved by designing

the software architecture of transmitter through three modules:

i) Micro-PC module. This module generates packets and communicates with the slave side of the USB dongle through host software.

ii) USB module at the RF-frontend. This module is responsible to transfer data packets generated from a higher layer to the RF module in respond to the request of the host software.

iii) RF module. In this module, the data packets are encapsulated with preamble, sync word, length and Cyclic Redundancy Check (CRC) and then modulated and finally transmitted over the air.

The performance of our development has been evaluated through various test scenarios which emphasize the validity of our result.

The rest of the paper is organized as follows: Section 2 provides an overview of our proposed transmitter regarding software and architecture requirements. Section 3 discusses the implementation in details and how the packet is designed and also how the transmission process performs. Experimental results through various tastings are described in section 4. Conclusion and future works are left for the section 5.

2. Transmitter design

Our proposed wireless transmitter composed of micro-PC (i.e., Raspberry Pi) and RF transmitter (i.e., CC2511F32, USB dongle from Texas instruments). This low cost, portable and low power device is capable of handling different RA protocols. Following sections discuss the design requirements of this transmitter.

2.1 Software requirement

We classify the software requirements in two categories: functional and non-functional requirements. Functional requirements represent the functions of the system. In our case, the wireless transmitter generates the packets by reading a file of any format (i.e., txt, bmp, etc.). Generated packets move from MAC layer to physical layer via the USB interface and finally the packets will be transmitted over the air. Non-functional requirements relate to the tools used for development of the software. In our case, the software development (IDE) and debugging have been done using IAR embedded work bench using embedded C programming language through incremental development methodology [11]. As mentioned before, the design of the whole software module is divided into three modules: USB module, radio module and micro-PC interface. The complete software development is based on incremental development. Therefore, each module is developed and tested in a series of versions. New developed version of each module is integrated with the old version. Finally the complete software is developed and validated through its assigned version number.

We consider whole system as an object that can be partitioned in to several smaller objects and layers. Some of these objects are reusable while others need to be modified according to the hardware specifications. In our case, the SoC (System-On-Chip) is CC2511F32 USB dongle which is based on SoC CC2510F32. The most architecture and register settings such as radio module, Direct Memory Access (DMA) and Analog to Digital Converter (ADC) are the same in both SoCs except for the extra USB module.

As shown in the figure 1, the software architecture has been divided in to three distinct layers. The outermost is the application layer which specifies the system behaviors and performs functional and user requirements. The next is the Hardware Abstraction Layer (HAL). In this layer various core functionalities have been defined. The HAL has been divided into two distinct parts: the common and the target specific part. The common part contains common software which can be portable for most of the targets (e.g., a radio module). The target specific part contains specific software to run on a particular hardware platform. It also contains functionality related to the user interface (e.g., Light Emitting Diode (LED) and Liquid Crystal Display (LCD) module or I/O interface, clock, USB module, etc.).

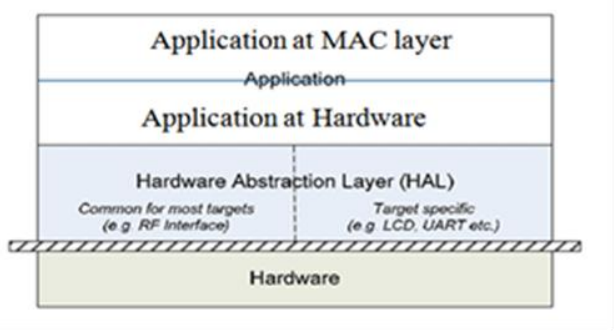


Fig. 1 Software architecture

The innermost layer is hardware dependent layer, which determines an appropriate action and needs to be taken for given set of inputs. These inputs drive the outputs to the desired state. The layer is pre and post-processed by the registers at the hardware end.

2.2 Hardware requirement

The hardware architecture of the wireless transmitter consists of two parts: micro-PC and RF transceiver. The micro-PC and RF-transceiver "CC2511F32" are the products of Raspberry Pi and Texas instruments respectively. The RF-transceiver is a low-cost 2.4 GHz SoC and is mainly used for low power wireless applications [12]. The micro-PC is a credit card sized, low power, affordable and easy to handle. The block diagram of the wireless transmitter is shown in figure 2.

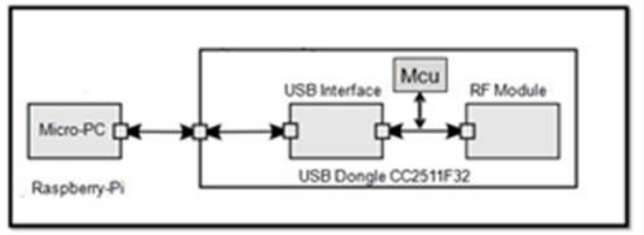


Fig. 2 Hardware architecture

2.3 USB frontend

The main objective of our task is to implement the USB interface between the host and RF module. Therefore, when the host sends a request to the USB controller, it has to respond based on the host request. The micro-PC acts as a host that initiates the communication. The USB dongle acts as slave device to the host. The communication by the USB dongle is done serially due to the supported drivers (e.g., Communication Device Class (CDC), Abstract Control Model (ACM) which emulate like serial device or virtual UART). After implementation of the USB interface, we need to implement the RF module to be able to transmit the packets generated from the micro-PC with the correct modulation and data rate. Therefore the whole task has been divided into two modules: The USB module and the RF module. In the following sections we will discuss the implementation details of USB modules and how it can be interfaced with the RF module and finally transmission of the packets over the air by RF module.

2.4 Radio model CC2511F32

Figure. 3 shows the block diagram of Radio model CC2511F32. By this device, the received RF signal is amplified by a Low Noise Amplifier (LNA) and is converted down in I/Q (in phase and quadrature phase signal) to Intermediate Frequency (IF). I and Q signals are digitized by ADCs. Later on Automatic Gain Control (AGC) and fine channel filtering and demodulating the received signal and packet synchronization are done digitally. In the transmitter side, synthesis of the RF frequency and the frequency synthesizer consist of an on-chip LC Voltage Controlled Oscillator (VCO) and a 90 degree phase shifter which generate I and Q Local Oscillator (LO) signals to down conversion mixers in receive mode. The high frequency crystal oscillator is used to generate reference frequency for frequency synthesizer and also for clocks for ADC and digital part.

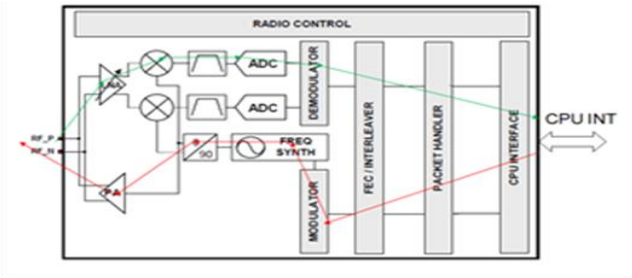


Fig. 3 Radio model CC2511F32 [7]

The digital baseband contains packet handling, channel configuration and data buffering. The SFR interface is used to access the data buffer via CPU. Control and status information are accessed via XDATA memory. In order to configure the radio module, there are a set of command strobes used by the CPU. These are single byte instructions in order to enable transmit and receive mode and also to enable and calibrate frequency synthesizer. These commands are as follows [7].

- STX: If in idle mode, perform calibration and enable transmit mode.
- SRX: If in idle mode, perform calibration and enable receive mode.
- SIDLE: Idle mode when no transmit or receive mode and the frequency synthesizer is OFF.
- SFXTXON: Enable and calibrate frequency synthesizer.
- SCAL: Calibrate the frequency synthesizer and turn off.
- SNOP: No operation.

The RF transceiver is based on the industry standard CC2500 with following characteristics:

- Operating frequency band of 2480 - 2483.5 MHz.
- Supports packet oriented system, on-chip for preamble detection, sync word detection, address check, variable and fixed packet length mode and automatic CRC check.
- Supports use of DMA. There is minimal intervention from CPU at high data rates.
- Supports programmable channel filter bandwidth.
- Supports three different modulation schemes: 2-Frequency Shift Key (FSK), MSK (Minimum Shift Key) and Gaussian Minimum Shift Key (GMSK).
- Optional automatic whitening and de-whitening of data.

- Programmable Carrier Sense (CS) indicator.
- Programmable preamble quality indicator and improved protection of sync word detection against random noise.
- Supports automatic clear channel assessment.
- Supports Link Quality Indicator (LQI).

The CC2511F32 has a built in state machine that can be used to switch between different operating modes. The radio switches to different states through command strobes or by internal events, such as TX_UNDERFLOW. The state of the radio can be figured out by reading the MARCSTATE status register. The radio has two active modes (i.e., STX and SRX). Writing these command strobes to the RFST register will initiate a TX or RX process. Whenever the radio enters TX mode, at first, a frequency check is done, then checks whether the radio is configured with High speed Crystal Oscillator (HSXOSC) with correct clock frequency and then switches to SIDLE (i.e., idle mode). Whenever the radio switches from SIDLE strobe to STX or SRX or vice versa, frequency synthesis or calibration is done. When transmit mode is activated, the chip will remain in Transmit (TX) state until the packet has been transmitted. After the packet has been transmitted, the radio changes the state. After transmit mode, the radio switches to receive mode through strobe command SRX. Figure. 4 shows the state diagram of the radio module.

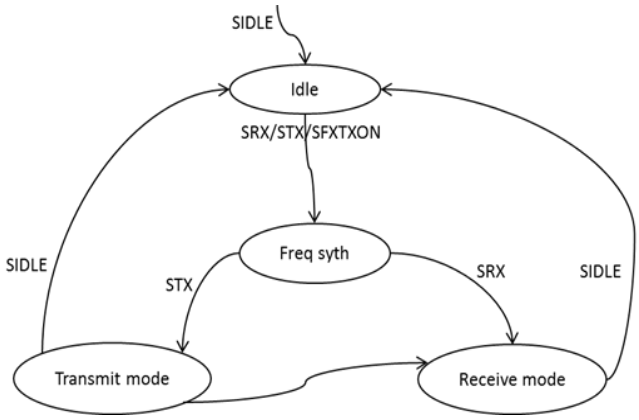


Fig. 4 State diagram of radio module [7]

2.5 Micro-PC

The micro-PC (i.e., Raspberry Pi) needs to be configured to control the USB based RF-frontend in order to initiate packet generation and transmission based on the developed RA MAC protocol for effective sharing of the medium. The task is to design a micro-PC based interface (i.e., how the host software which communicates with the USB based RF-frontend generates packets by reading a file of any size that may be in a txt, bmp, etc. format). After the file is read, it is encapsulated with a header and then transmitted via USB interface to RF module. The following section begins with an introduction to Raspberry Pi as well as developed method to implement micro-PC based interface. Further, we will discuss packet generation along with its testing procedure.

Raspberry Pi

The Raspberry Pi is a credit-card sized single board device developed by the Raspberry Pi foundation [13][17]. It is based on Broadcom 700MHz SoC and is the main central module for controlling the whole transmitter which contains a 32bit ARM1176JZF-S with 700 MHz RISC processor and a Video Core IV GPU. As shown in figure. 5, the Raspberry Pi is composed of a processing unit, memory, power supply, HDMI output, Ethernet port, USB ports and other interfaces. The main reason for choosing it as a micro-PC is its low cost, small in size, low power consumption and its support by Linux based operating system for developing the host software as well as developing various random access protocols at the MAC layer. The micro-PC acts as an intelligent system in which all the events are generated and the firmware at the USB based RF-frontend will respond based on the request of the host. A host is a PC which contains a host-controller and software.

As mentioned before, this project has hardware and software requirements. The hardware requirements have been met by configuring the micro-PC (i.e., Raspberry Pi) as a central controlling device for the wireless transmitter. The Raspberry Pi is compatible with USB 2.0 high speed port so that the USB based RF- frontend can be interfaced with a micro-PC. As software requirements, there is a Linux based operating system in Raspberry Pi. We also used a gcc/g++ compiler in order to compile the developed software under Linux. Regarding the USB based RF-frontend, the driver at the USB based RF-frontend needs to be compatible with Linux (i.e., CDC ACM). In order to develop the host software at the micro-PC to communicate with the USB dongle, we used Libusb v.1.0 [14]. Libusb is an open source library and an API platform to develop software in order to communicate with the USB peripheral. This API supports all kinds of USB transfers such as bulk, interrupt, control and isochronous. This API supports synchronous as well as asynchronous transmission.

Here, the micro-PC acts an intelligence system in which different RA protocols are developed and it is also responsible for generating the packets and establishing communication between the micro-PC and the USB dongle for wireless transmission.

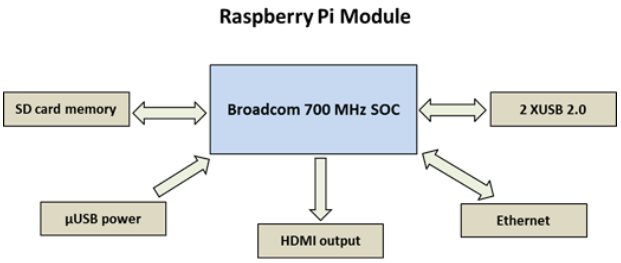


Fig. 5 Raspberry Pi module block diagram

As shown in Figure. 6, we describe how Libusb has been configured. The process begins with *Libusb initialization* (i.e., this indicates the start of session and this function must be called before any other function of Libusb). *Get device* list function will list all the devices connected to the micro-PC and *Get device descriptor* function will read the structures related to device that is located in the flash of a device like VID, PID, etc. The *Open* function opens the device, (i.e., in this case USB dongle) and returns the handle of the device. Further, the *conditional function* looks whether the kernel driver is active or not. If it is *active*, then detach the event, otherwise, the device can *claim* the interface. Once the device claims the interface, it starts communication with the USB dongle by sending control data using a *control* transfer. This function is used to configure the device such as enumeration. For example, in this case, the micro-PC sends control signals such as CTS, RTS, DTR, baud rate, data bits, stop bit and parity bit in order to establish communication between the host and USB dongle. Since the USB dongle has a CDC ACM driver, it emulates a virtual UART. Using bulk transfer mode, the micro-PC is able to transmit the dummy data (e.g., 64 byte) packets to the USB dongle using an OUT endpoint address. The dongle transmits the packets over the air and on the receiver end, the Evaluation Board (EB) which acts as a receiver, along with packet sniffer tool capture the packets from the micro-PC based wireless transmitter. Finally, the interface is released and Libusb session is *closed*.

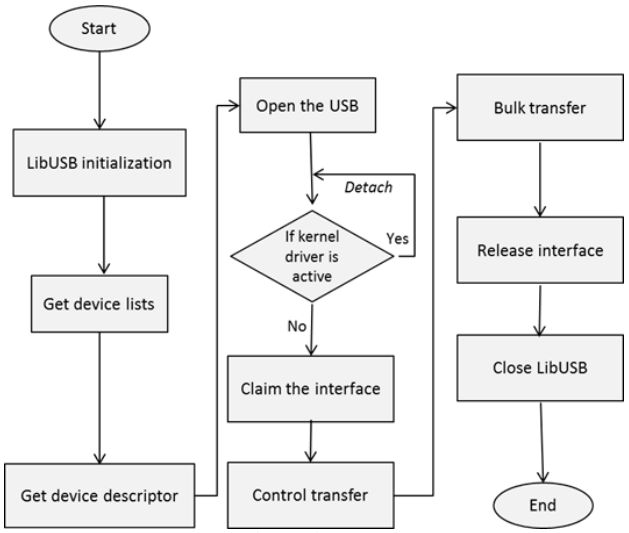


Fig. 6 Host software flow chart using Libusb

3. Implementation

This section explains the design of packet format generated by micro-PC and then the transmission operation in details.

3.1 Packet transmission and reception

The packet has been configured based on the following format shown in Figure. 7.

- Programmable preamble; 8-24 byte
- Programmable synchronization word16 or 32 bits
- Programmable or constant length byte
- Address (optional).
- Payload.
- CRC (optional).

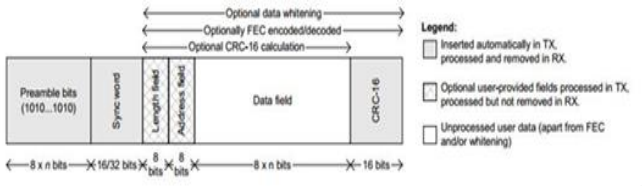


Fig. 7 Packet format [7]

The CC2511F32 has built in hardware support for packet oriented radio protocol. Preamble can be modified from 2-24 bytes and synchronization (Sync word) from 2 to 4 bytes. The length byte is optional and is enabled when the variable packet length mode is selected. Note that in fixed length packet transfer mode, the first byte in the payload will be the destination address if it is enabled, otherwise it will be payload. The maximum packet length is 0-255 bytes and an optional 2 byte CRC which is calculated over the data if it is enabled [7].

The CC2511F32 USB dongle supports three kinds of modulation schemes (i.e., MSK, 2-FSK, Gaussian frequency-shift keying (GFSK)). In this model, we selected MSK because it is similar to GMSK, which is the modulation used for the Automatic Identification System (AIS). MSK is a subclass of Continuous Phase Frequency Shift Key (CPSFK) and the MSK spectrum has no discrete components unlike other modulation. The MSK spectrum is wider than QPSK (Quadrature Phase Shift Key) but the side lobes fall much faster than QPSK. MSK encodes each bit as half sinusoidal and because of this feature, it has a constant envelop, compact spectrum, good error performance which it is mostly used in wireless systems [8][9]. Table 1 shows different modulation and data rates the USB dongle supports.

Table 1: Modulation

Modulation	***Min***	***Max***	***Unit***
MSK	26	500	kBaud
2-FSK	1.2	500	kBaud
GFSK	1.2	250	kBaud

The following procedure explains the packet transmission and reception.

In order to transfer packet, the data has to be written into RF data register (RFD). In receive mode, the data has to be read from same register. The RFD register has 1 byte FIFO in TX mode if the number of bytes written in RFD register is less than what it has assigned to transmit. Then the radio will enter into TX_UNDERFLOW state. RFIF.IRQ_TXUVF and RFIF.IRQ_DONE flags are set and when a byte is not read from the data register and the next byte is ready to be received, then TX_OVERFLOW flag is set. If no data is written in the RFD register and a STX strobe is issued, after the assertion of the RFTXRXIF flag, the radio will start sending the preamble without going into TX_UNDERFLOW state. A temporary FIFO is created in memory in TX/RX (i.e., TX FIFO and RX FIFO).

In transmit mode, the DMA transfers data from TX FIFO to RFD register and optional length field if variable packet length is enabled. If fixed packet length is enabled, then the first byte of payload is the destination address. The modulator sends the number of preamble and then 2 or 4 bytes of sync word, data content in the payload and 2 bytes CRC which is calculated over the payload. Once receiving the data, it moves from RFD to RX FIFO [7].
In receive mode, the demodulator and packet handler will look for the correct number of bytes of preamble and sync word. When found, the first byte is read. If variable packet length mode is used, then the first byte is the length byte and the packet handler stores the value as a packet length and based on the length of the byte, it checks the received data. If it's programmed as fixed length, then the payload is read and optional CRC check will be done.

3.2 Packet design and transmission process

This section gives a brief overview on how we have designed the packet generated by micro-PC and then the transmission operation in details.

3.2.1 Packet design

The detailed design of the packet format is represented in Figure. 8.

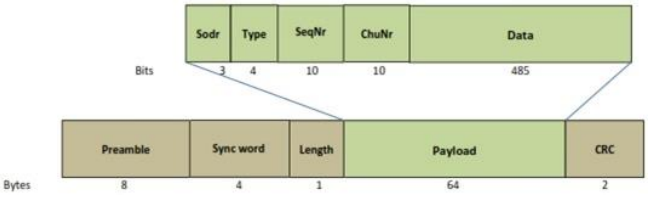


Fig. 8 Designed packet format

The data packets sent by the transmitter are generated by micro-PC. The payload of the data packet can be a portion of an image or text. The header of the packet contains:

- Source address (Sodr); distinguishes different transmitters.
- Type field; describes the data type of the payload.
- Sequence number (SeqNr); represents the portion of the file that is placed in the payload.
- Number of chunks (ChNr); represents the number of packets needed to send entire file.

Once the packet is generated, the MAC packet is moved via the USB interface to the RF-frontend using the control and bulk transfer mode of USB protocol and then transmitted over the air.

3.2.2 Transmission process

The transmission operation has been shown in details in Figure. 9. At the micro-PC end, the packet is generated with above defined format. The header is composed of 3 bits of source address, 4 bits of file type, 10 bits of sequence number and a payload of 64 bytes which is generated by MAC layer. In programming aspect, generally, two structures have been used. One defines the header of the packet and the other encapsulates the header and payload. The union includes both structures which is simply a whole packet. For packet transmission, a buffer must be allocated to load the complete file. The serial port has been configured with a baud rate of 38400 bps and has been provided with 8 data bits, no parity and 1 stop bit. Handshake signals have been enabled before starting communication. The packet header and payload are

autonomously filled once the program loop starts based on the total number of chunks which calculated through file size and payload. For example, to transmit 600 bytes (need 10 chunks, each of size 60 bytes) file, the sequence number is calculated based on the total number of chunks and the number of times the loop runs. Source ID is fixed and assigns a number to file type. The packets must be written to the USB port with a chunk size of 60 bytes along with 4 bytes of header information. Once the host sends the control signals, it is ready to transmit packets. Based on the preceding events, the firmware on the dongle will start responding to the host. The USB controller continuously waits for a packet and once it receives a 64 byte packet or less, it transfers the data to the endpoint of the FIFO. The CPU on the USB dongle reads the FIFO data and moves it to a TX buffer. A packet array has been created from the data held in the TX buffer. The DMA moves the packet array to RFD data register where the STX strobe is initiated. Once the strobe is initiated, the packet is transmitted over the air and the radio goes in RX mode.

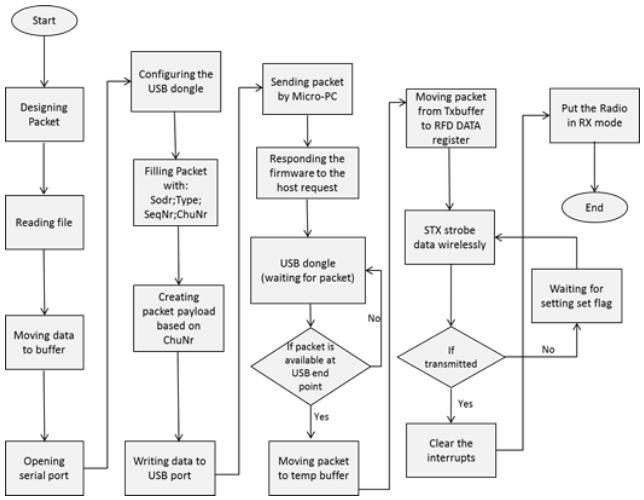


Fig. 9 Transmission process

4. Software testing and validation

This section discusses how the transmitter is tested along with the test bench setup and the software tools. We have developed the host software to communicate with the USB dongle. It is able to transmit data from the micro-PC towards the endpoint of the USB dongle. For example, by sending 64 bytes of dummy data, the dongle acts as a slave device and the firmware is able to respond based on the request of the host. During the test bench setup, the micro-PC based RF-frontend operates as a transmitter along with the packet sniffer tool to capture the radio packets from the micro-PC based RF-fronted. Further the test has been continued with the Smart-RF CC2510CC2511DK development kit from Texas instruments with CC2510F32 SoC which operates as a receiver. The CC2510F32 is a 2.4 GHz SoC which is based on a high performance leading transceiver CC2500. The CC2510F32 and CC2511F32 USB dongle have similar architecture except for clock frequency with 24 - 27 MHz for the CC2510 and 48 MHz for the CC2511. The EB has been configured to hold the characteristics of transmitter. In order to begin sniffing of the transmitter, the Smart-RF EB with CC2510F32 SoC is flashed with firmware from Texas instruments (i.e., sniffer_fw_ccxx10_usart0_alt1.hex) through specifications such as transmitting frequency, data rate and modulation. To sniff the radio packets of the capturing device, the radio configuration tab in the packet sniffer tool needs to have the configuration file of capturing radio which is generated using the Smart-RF studio. The procedures are as follows:

- Flash the transmitter HEX file in the USB dongle using Smart-RF flash programmer.
- Plug the USB dongle into the micro-PC USB port and plug the CC2510F32 SoC EB to the PC.
- Open and run the packet sniffer tool and configure the EB with radio configuration file which is generated from Smart-RF studio.
- Run the host software in the micro-PC. It generates the packets with header and payload by reading the file of any size at higher layer.
- Push the packet to a lower layer and transmit it over the air. On the other end, the receiver receives the packets and then towards sniffing tool.

The radio module has been configured with a base frequency of 2480 MHz, carrier frequency of 2479 MHz with data rate of 500 KBaud, received filter bandwidth of 750 KHz with MSK modulation, zero phase transition time, channel spacing of 199.951172, 0 channel number and with 0 dbm TX power. We have performed three test cases as follows.

4.1 First test case

As a first test case, the radio module has been tested. In the radio module, the packets are generated within the USB dongle with a certain packet format such as 1 byte of packet length, 2 bytes of TX ID and 4 bytes of sequence number and dummy payload. The generated packets have been transmitted by the dongle successfully over the air and on the other end, the EB receives and captures the packets using the packet sniffer tool.

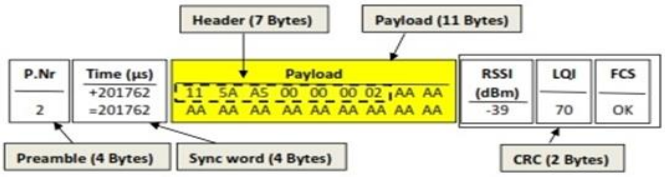


Fig. 10 Sample packet format of radio module

The packet format and sniffer data (captured packets) are shown in figure. 10 and 11 respectively. The First byte of payload represents the length byte and the next two bytes represent TX ID to distinguish among different transmitter at receiver end. The next 4 bytes represents as a sequence number to identify which packets have been received successfully and the reaming data is payload.

P.Nr	Time (µs)	Payload	RSSI (dBm)	LQI	FCS
1	+0 =0	11 5A A5 00 00 00 01 AA AA AA AA AA AA AA AA AA AA	-39	49	OK
2	+201762 =201762	11 5A A5 00 00 00 02 AA AA AA AA AA AA AA AA AA AA	-39	70	OK
3	+201761 =403523	11 5A A5 00 00 00 03 AA AA AA AA AA AA AA AA AA AA	-39	70	OK
4	+201761 =605284	11 5A A5 00 00 00 04 AA AA AA AA AA AA AA AA AA AA	-39	65	OK
5	+201761 =807045	11 5A A5 00 00 00 05 AA AA AA AA AA AA AA AA AA AA	-39	80	OK
6	+201761 =1008806	11 5A A5 00 00 00 06 AA AA AA AA AA AA AA AA AA AA	-39	78	OK
7	+201761 =1210567	11 5A A5 00 00 00 07 AA AA AA AA AA AA AA AA AA AA	-39	100	OK

Fig. 11 Packet sniffer data of radio module

4.2 Second test case

The second test case has been done over the micro-PC based RF transmitter module using Libusb 1.0 API. The setup and the radio configuration remain same as previous test with some more configurations in the host software at micro-PC. Using Libusb API 1.0, the communication has been established between the micro-PC and RF-frontend. Dummy packets of 64 bytes have been generated at the micro-PC. Upon generation of these packets from the higher layer, they are transferred using control and bulk transfer mode via the USB interface to the physical layer. When packets are received at the RF-frontend, they are transmitted over the air. On the other end, packets are received and captured by the EB using packet sniffer tool. The packet format and captured packets are shown in figure. 12 and 13 respectively. Here, the first byte is length byte and remaining is payload.

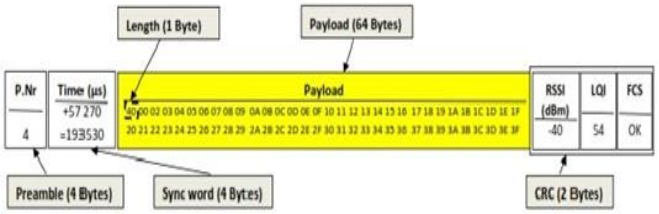


Fig. 12 Sample packet format of micro-PC

P.Nr	Time (µs)	Payload	RSSI (dBm)	LQI	FCS
1	+0 =0	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	82	OK
2	+69199 =69199	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	44	OK
3	+67061 =136260	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	54	OK
4	+57270 =193530	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	54	OK
5	+55155 =248685	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	68	OK
6	+62088 =310773	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	64	OK
7	+54099 =364872	40 00 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F 10 11 12 13 14 15 16 17 18 19 1A 1B 1C 1D 1E 1F 20 21 22 23 24 25 26 27 28 29 2A 2B 2C 2D 2E 2F 30 31 32 33 34 35 36 37 38 39 3A 3B 3C 3D 3E 3F	-40	80	OK

Fig. 13 Packet sniffer data of micro-PC

4.3 Third test case

The final test case has been carried out through micro-PC based RF transmitter using POSIX terminal interface as host software [15]. The test setup remains the same as before, but the dongle is flashed with the firmware of a USB RF-transmitter. This host software has some features such as the ability to generate the packets by reading a file of any size and encapsulates them with the header. The maximum size of the packet that can be transmitted is 64 bytes due to the limitation at the USB RF- frontend (i.e., the USB diver CDC ACM supports maximum packet size of 64 bytes).

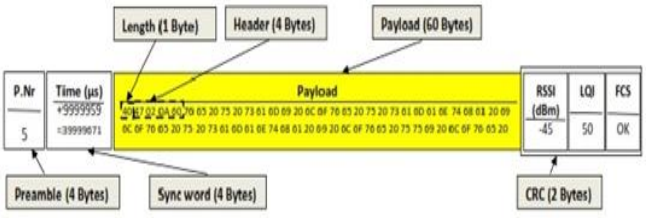


Fig. 14 Sample packet format of micro-PC using POSIX terminal

To better assess the impact of packet size, in this test case, we have examined small and large packet sizes with same radio configuration as mentioned above. At the first step, the host software reads small size file of 311 bytes. It starts by reading a complete file and allocating a buffer according to the size of the file. The reason is that the host software cannot send a complete file due to limitations at the USB dongle (i.e., maximum size is 64 byte for CDC ACM). Therefore, the file is divided into 64 bytes chunks with 4 bytes of header and 60 bytes of payload. Note that, the total number of chunks is calculated using the size of the file and maximum size of the payload that can be transmitted via the USB RF-frontend. For example, to

transmit 311 bytes, the maximum payload that can be transmitted is 60 bytes and we require 6 packets to complete the whole transfer. Figure. 14 shows the packet format and figure. 15 shows the complete transfer of 311 bytes.

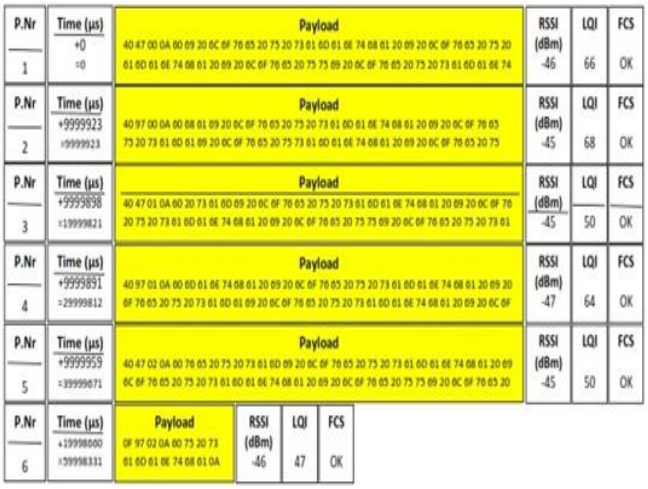


Fig. 15 Packet sniffer data of micro-PC using POSIX terminal for 311 bytes

The second step of the test has been done by reading a file with 416 bytes size and transmitting it via USB interface to RF-frontend and then transmitting it wirelessly as shown in figure. 16.

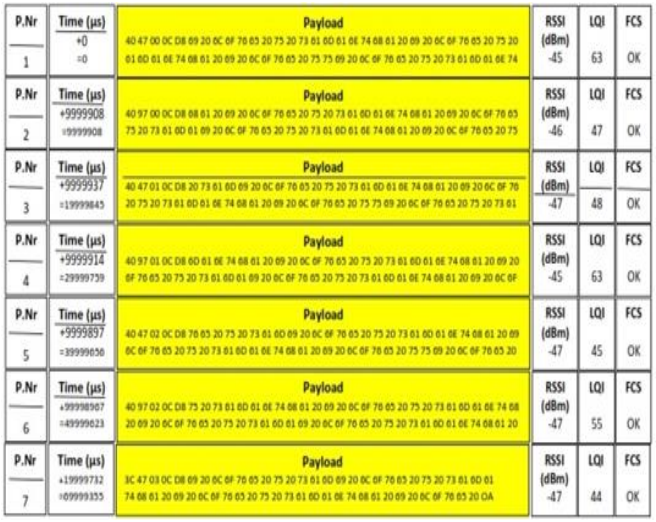


Fig. 16 Packet sniffer data of micro-PC using POSIX terminal for 416 bytes

5. Conclusions

In this work, we have designed and developed a portable random access wireless transmitter composed by micro-PC and a radio frontend working in the WiFi frequency bands to support different random access protocol. To achieve this RA solution, we needed a deep understanding of proposed hardware architecture to implement different modules such as USB, radio and micro-PC. At the beginning, the packets at the micro-PC side have been successfully generated. Then an appropriate USB link between the micro-PC and the RF-frontend at the USB dongle has been implemented. By setup configuration, the RF frontend is able to transmit and receive wireless data using the evaluation board. The complete transmitter with the micro-PC based RF-frontend along with the CC2510F32 EB which emulates a receiver, have been finally tested with a packet sniffer tool. As a result of this project, the generated radio packets at the micro-PC side have been transmitted over the air successfully without any packet loss.

As for the future work, the transmission blocks at the developed micro-PC end, such as encoding, filtering, etc., has to be implemented and tested using various random access protocols.

Acknowledgments

Part of this work has been carried out at the Institute of Communications and Navigation of the German Aerospace Center (DLR) in Oberpfaffenhofen, Germany. The Authors would like to thank Dr. Andrea Munari and Mr. Federico Clazzer who supervised this work at DLR Company for the resources and technical support [16].

References

[1] P. Pal, "Medium access control (mac) techniques", IIT, Kharagpur, available at "http://nptel.ac.in/courses/106105080/pdf/M5L2.pdf", accessed on 13/11/2015.
[2] R. Rom, M. Sidi, "Multiple access protocols: performance and analysis", available at "http://books.google.de/books?id=yAVTAAAAMAAJ", accessed on 13/11/2015.
[3] N. Abramson, "The throughput of packet broadcasting channels", IEEE Transactions on communication, vol. 25, no. 1, pp. 117–128, 1977.
[4] B. Chen, L. Tong, "Traffic-aided multiuser detection for random- access cdma networks", IEEE Transactions on Signal Processing, vol. 49, no. 7, pp. 1570–1580, 2001.
[5] H. Lee, "Random access schemes for multichannel communication and multi-packet reception in wireless networks", Ph.D. dissertation, University of Adelaide, 2011.
[6] Rashid Hassani, Ganesh Chavan, Peter Luksch, "Optimization of Communication in MPI-Based Clusters", In proceedings of the international conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC 2014), pp. 143-149, 2014.
[7] "Low-Power SoC (System-on-Chip) with MCU", Texas Instruments, available at

"http://www.ti.com/lit/ds/swrs055g/swrs055g.pdf", accessed on 13/11/2015.
[8] S. Pasupathy, "Minimum shift keying: A spectral efficient modulation", IEEE Communications magazine, available at "http://www.q.hscott.net/reads/QMSK.pdf", accessed on 13/11/2015.
[9] Rashid Hassani, and Peter Luksch, "Optimizing Bandwidth by Employing MPLS AToM with QoS Support", In proceedings of the IEEE NAS 2012, pp. 104-108, 2012.
[10] Lei Zheng, Lin Cai, "AFDA: Asynchronous Flipped Diversity ALOHA for Emerging Wireless Networks with Long and Heterogeneous Delay", IEEE Transactions on Emerging Topics in Computing, pp. 1-11, 2014
[11] "IAR embedded workbench", Debugging using the IAR C-spy debugger, available at "http://supp.iar.com/FilesPublic/UPDINFO/005832/arm/doc /infocenter/ tutor_debugging.ENU.html", accessed on 13/11/2015.
[12] "Silicon Labs", USB overview, available at "http://www.silabs.com/Support% 20Documents/Software/USB_Overview.pdf", accessed on 13/11/2015.
[13] V. Sathish Kumar, G. Senthilkumar, K. Gopalakrishna, "Embedded image capturing system using raspberry pi system", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 3, no. 2, pp. 213-215, 2014.
[14] "libusb", available at "http://libusb.sourceforge.net/api-1.0/", accessed on 13/11/2015.
[15] Michael R. Sweet, "Serial Programming Guide for POSIX Operating Systems", available at "http://www.netzmafia.de/skripten/ hardware/Seriell/SerialPort_Programming_c.pdf", accessed on 13/11/2015.
[16] "University of Rostock", available at "http://wwwvhr.informatik.uni-rostock.de/projekt_bachelor_und_master_arbeiten/vhr_diplomarbeiten/", accessed on 13/11/2015.
[17] Vujovic. V, Bosnia Herzegovina, Maksimovic. M, " Raspberry Pi as a Wireless Sensor node: Performances and constraints", Information and Communication Technology, pp. 1013-1018, 2014
[18] Cedomir Stefanovic, Petar Popovski, "ALOHA Random Access that Operates as a Rateless Code", IEEE Transactions on Communications, vol. 61, no.11, pp. 4653-4662, 2013.
[19] Minh Hanh Ngo, Vikram Krishnamurthy, Lang Tong, "Optimal Channel-Aware ALOHA Protocol for Random Access in WLANs With Multipacket Reception and Decentralized Channel State Information", IEEE Transactions on signal processing, vol. 56, no.6, pp. 2575-2588, 2008.
[20] Rashid Hassani, Shiv. R. P. N Amgoth, Peter Luksch, "Efficient Consolidation of Virtual Machines for HPC Applications in Cloud", International Journal of Intelligent Information Processing, vol. 5, no. 4, pp. 19-26, 2015.

Rashid Hassani received his BE degree in Information Technology from VTU, India. He graduated with M.Sc. degree in Computer System Engineering emphasized in Embedded and Cooperating Systems from Halmstad University, Sweden. He is currently a research assistant and academic teacher in the department of computer science at University of Rostock, Germany. His research interests include parallel and high performance computing, Cloud computing and networking protocols.

Prabhu Gudapusetty received his M.Sc. degree in Computational Engineering at university of Rostock.

Peter Luksch received his Ph.D. degree in Parallel Discrete Event Simulation on Distributed Memory Multiprocessors from Technische Universität München, Germany. He finished his Postdoctoral Lecture Qualification (Habilitation) in Increased Productivity in Computational Prototyping with the Help of Parallel and Distributed Computing. Currently, Prof. Dr. rer. nat. habil Peter Luksch is head of the department of Distributed High Performance Computing at University of Rostock, Germany.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

An Intelligent System based on Fuzzy Inference System to prophesy the brutality of Cardio Vascular Disease

Sivagowry S[1], Durairaj M[2]

[1] Department of Computer Science, Engineering and Technology, Bharathidasan University, Tiruchirapalli, Tamilnadu 620024, India
sivagowry87@gmail.com

[2] Department of Computer Science, Engineering and Technology, Bharathidasan University, Tiruchirapalli, Tamilnadu 620024, India
durairaj.bdu@gmail.com

Abstract

To unravel hidden relationships and diagnose diseases efficiently, Data Mining along with Soft Computing Techniques are used in several researches. Cardio Vascular Disease is a condition which leads to severe disability and death. Since the diagnosis involves vague symptoms and tedious procedures, diagnosis is usually time-consuming and erroneous. For the healthier analysis and treatment of heart disease based on brutality, an Intellectual, accurate and proficient investigative system is needed. For diagnosing heart disease with improved effectiveness, an Intelligent Fuzzy Inference System is needed. This paper illustrates how Fuzzy Inference System is used to envisage the severity of disease by constructing an effective Fuzzy Rule Base. It is also proved that a precision of 95.23% is obtained when Fuzzy System is used in severity prediction

Keywords: *Fuzzy Inference System, Fuzzy Rule Base, Severity prediction, Heart disease, Mamdani fuzzy system***.**

1. Introduction

The role of Data Mining in health care data is massive. The human decision making is optimal, but it is poor when the amount of data to be classified is huge. The enormous stress and overwork load resulted in poor / inaccurate decision making which may lead to disastrous consequences and cannot be allowed in medical field. The most exorbitant and harmful mistake is performing decision making process based on improper information acquired from medical data [1]. Institute of Medicine estimated that the effect of medical error accounts for about $17 to $29 billion, which is not declined since then. Medical history data, which comprise of number of tests and previous examination cycles, is essential to diagnose and devise future course of treatments on particular disease. It is conceivable to increase the benefit of Data mining [2], [3], [4] in health care by employing it as an intellectual symptomatic tool [5]. The researchers in the medical field have prospered in categorizing and prophesying the syndrome with the encouragement of Data mining techniques [6]. Association rules of Data Mining have been significantly used in health care data prediction [7], [8], [9, and 10]. The eventual goal of knowledge discovery is to identify factors which tend to improve the quality and effectiveness of health care system.

2. Heart Disease

The rise of health care cost is one of the universally confronting problems. The therapeutic term for Heart Disease / Heart Attack is Myocardial Infarction (MI) or Acute Myocardial Infarction (AMI). Heart attack emerges when there is indiscretion in the flow of blood and bruised heart muscles due to inadequate oxygen supply [11]. Jeopardy factors for Myocardial Infarction include smoking, high blood pressure, cholesterol, Diabetes, Family history, etc.,. Cardio Vascular Disease (CVD) clinical guidelines spotlight on the management of single risk factors [12, 13, and 14]. In majority of cases, it doesn't work since risk factors crop up in clusters, that is, the presence of single risk factor which indicates the presence of other risk factors too. It is apparent that the presence of multiple risk factors increases the sternness of CVD. World Health Organization (WHO) in the year 2008 testified that 30% of total global bereavements are due to Cardio Vascular Disease (CVD). WHO has announced India as global CVD capital [7]. Rapid urbanization and industrialization have led to the major cause of demise in India due to CVD [8]. It is estimated that numeral of CVD patients will increase from 300 million to 600 million by 2020 [15, 16]. By 2030, almost 25 million people will die from CVDs, mainly from heart disease and stroke [17], [18], [19]. These are projected to remain the CVD is the single leading cause of death. The CVD is also expected to be the leading cause of deaths in developing countries due to changes in lifestyle, work culture and food habits. Hence, more careful and efficient methods of diagnosing cardiac diseases and periodic examination are of high importance [11] [10] [20] to take preventive care.

The paper is organized as follows: Section III gives a magnitude about the existing literature work regarding

Fuzzy logic in Heart disease prediction. Section IV describes the data set collected for experimentation. Section V describes how the Fuzzy Inference System evaluates the severity of heart disease based on Fuzzy Rule base. Section VI discusses the results. Section VII concludes the paper.

3. Review of literature based on Fuzzy Logic in Heart Disease Analysis

A new Particle Swarm Optimization based Fuzzy expert system was proposed which involves four stages [21]. Nearest hot deck imputation is used to remove the missing data. A Fuzzy Expert System is generated based on set of rules. So the system can used to provide interpretations for decisions. The Fuzzy Expert System which can be developed by using the set of rules and tuned Member Functions improves the accuracy. Mamdani fuzzy inference system is used for Fuzzification of crisp sets. The Centre of Gravity (COG) method is employed for defuzzification process. 93.27% of accuracy was obtained by using the Fuzzy Expert System.

Existing system was studied and a disease prediction system which is based on fuzzy is proposed. The accuracy provided by this system is more than 90% [22]. A neuro fuzzy network was designed to identify and classify the coronary artery disease by using MATLAB [23]. Sugeno-based fuzzy expert system is used. The sensitivity and specificity obtained are 1 and 0.88 respectively. This shows that the network have an acceptable degree of accuracy.

A data mining technique which is based on rough set theory and fuzzy logic is proposed. It has two phases which includes clustering and classification. Clustering is based on rough set theory and fuzzy are used for classifying the clusters obtained by using the rough set theory. Complexity in generating the rules based on Fuzzy logic is reduced since rough set theory is used prior to fuzzy logic. MATLAB is used for implementation [24]. Sensitivity, specificity and accuracy are the parameters which were used to measure the performance of the proposed system. From the result, the data set of Switzerland database provides better results than other. The proposed method can also used to deal with uncertainty problem.

Advanced fuzzy resolution mechanism [25] was conducted for diagnosing the heart disease. MATLAB Fuzzy Logic Tool box is used to generate the rules. It has five layers, and each layer has its own node. Accuracy is used as a metric to compare the working of proposed algorithm with existing methods. Accuracy obtained is 94.11% which is higher when compared with other methods.

Two systems were developed to diagnosis heart disease based on MATLAB [26]. Multi Layer Perceptron (MLP) network is the base for developing the first system. Second system was developed based on Adaptive Neuro Fuzzy Inference System (ANFIS). 80% of data was used for training and 20% is used for testing. The accuracy obtained while training by using ANFIS was 100% whereas by using MLP it was 90.74%. But, while testing, MLP outperforms ANFIS in accuracy

4. Data set description

The Data set used for experimentation is taken from Data mining repository of the University of California, Irvine (UCI). Data of Cleveland Data set, Hungary Data set, Switzerland Data set, Long beach and Statlog Data set are collected. Cleveland, Hungary, Switzerland and Va Long Beach data set contain 76 attributes. Among all the 76 attributes, 14 attributes are taken for experimentation, since all the published experiments refer to using subset of 14 attributes. Researchers in the medical domain mostly use Cleveland data set and Statlog data set for testing purpose. This is because all the other data set has more number of missing values than Cleveland data set [27]. The Table 1 describes the data attributes used in this work.

Among the 13 attributes, only 5 attributes are chosen by using the knowledge of Intelligent Hybrid Quick Reduct Particle Swarm Optimization Algorithm [28] which does not affect the accuracy in classification and prediction. The chosen five attributes are cp, exang, slope, ca and thal

5. Fuzzy Logic in Severity Prediction

The following is the algorithm which is used to construct the rule base Fuzzy Inference System for severity prediction.

Fuzzy Logic Algorithmic steps:
Input:
cp, exang, slope, ca, thal
Output:
Heart Disease Severity- Low, Mild, Severe
Begin
Input crisp value
Set Mamdani fuzzy model, with fuzzy if-then rules.
Assign fuzzy member for the input of each variables.
Output variable is predicted by using the input.
Generate rules

Calculate Membership function by using Guassian Membership function given in equation (1)

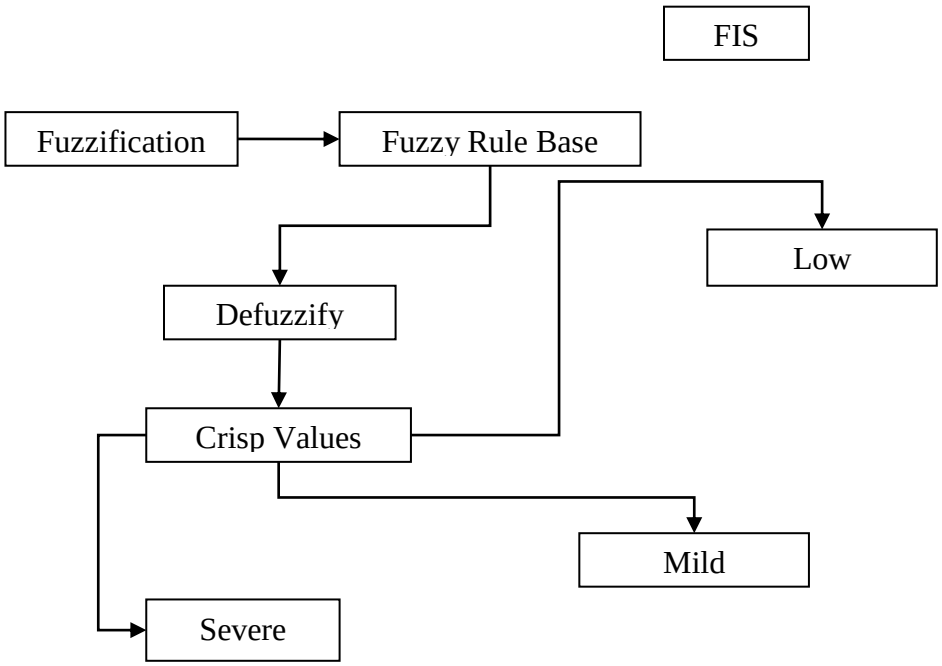


Figure 1 : Fuzzy Inference System for Severity Prediction

$$Guassian\ Membership\ Function = \mu(x,a,b) = e^{\frac{-(x-b)^2}{2a^2}} \quad .....(1)$$

Fuzzy rule base Construction.
Defuzzification of the fuzzy value to get the final crisp output. The Centroid method is used for defuzzification.

$$Centroid = ZCOA = \frac{\int_Z \gamma_A(z) z dz}{\int_z \gamma_A(z) dz} \quad .....(2)$$

End

The Figure 1 describes how the Fuzzy Inference System is designed for predicting the severity of Cardio Vascular Disease.

The five inputs cp, slope, ca, exang, thal and the output severity is converted into fuzzy variables as follows:

The input chest pain (cp) has four types namely typical, atypical, non-angina and asymptomatic. It is assigned the fuzzy values as 1,2.3 and 4. The figure 2 shows the membership function plot of the input variable cp

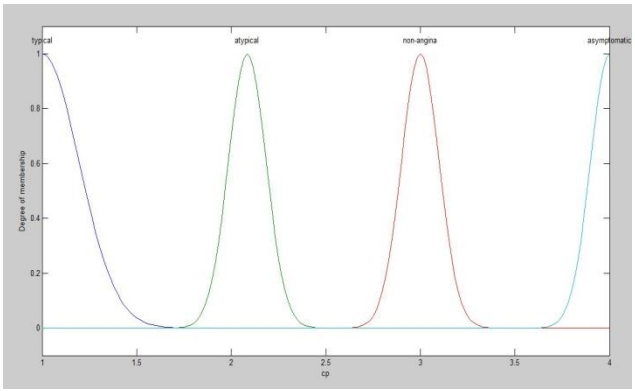


Figure 2: Membership function plot for the input variable cp

The next input variable exercise induced angina (exang) has two fields yes and no. It has 1 for yes and 0 for no. The field slope has three fields as shown in Table 2. The Table 2 shows the fuzzy value assigned for the crisp input for the input field slope

Table 1: Description of the data attributes

No	*Name*	*Description*
1	Age	Age in Years
2	Sex	1=male, 0=female
3	cp	Chest pain type(1 = typical angina, 2 =atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	Trestbps	Resting blood sugar(in mm Hg on admission to hospital)
5	chol	Serum cholesterol in mg/dl
6	fbs	Fasting blood sugar>120 mg/dl(1= true, 0=false)
7	Restecg	Resting electrocardiographic results(0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1=upsloping, 2=flat, 3= downsloping)
12	ca	Number of major vessels colored by fluoroscopy
13	thal	3= normal, 6=fixed defect, 7= reversible defect
14	Num	Class(0=healthy, 1=have heart disease)

Table 2: Fuzzy value for the slope

Upsloping	1
flat	2
downsloping	3

The number of major vessels colored by fluoroscopy is indicated by the input variable ca The Gaussian membership function is used to plot the membership value for all the fields. The final input variable is thal which is the size and location of injured muscle after the heart attack. The Fuzzy value for each of the value is given in Table 3. The value 3,6 and 7 indicates normal, fixed defect and reversible defect respectively.

Table 3: Fuzzy value for thal

normal	3
fixed defect	6
reversible defect	7

The figure 3 shows the Fuzzy Inference System designed for the extraction of fuzzy rule in predicting the severity of Heart Disease. The Figure 9 shows the description of the FIS designed using Fuzzy logic for severity prediction. After designing the FIS , the rule base is constructed by using the if-then rules. Number of rules framed for the system is 63. The Mamdani system is used for designing the FIS. Because the system has widespread application than the Sugeno model and it is more suitable system for the human input. So, the Mamdani system is used for designing the FIS for predicting the severity of the Heart disease.

Name = **Pref**
Type = **Mamdani**
NumInputs = **5**
InLabels = **cp**
exang
slope
ca
thal
NumOutputs = **1**
OutLabels = **Severity**
NumRules = **63**
AndMethod = **min**
OrMethod = **max**
ImpMethod = **min**
AggMethod = **max**
DefuzzMethod = **centroid**

Figure 4: Description of FIS for prediction

Fuzzy rule base is constructed using the reduced attributes. Rule viewer shows the strength of each rule and the surface viewer plot the output with respect to every input. The figure 5 shows the rule viewer of the prediction system designed.

6. Results and Discussion

The defuzzified value obtained by using centroid method is compared with original output and manual output calculation. The Table 4 shows the comparison of each value obtained. In each case, it was observed that there is no much difference in the output observed which supports the use of Fuzzy Logic in severity diagnosis. The accuracy in severity prediction has been increased. It was up to 95.86% when using Fuzzy LogicWhile using the Fuzzy logic the whole prediction level has been increased. The figure 6 shows the increase in accuracy when Fuzzy Logic is employed. The whole FIS performs well in predicting the severity level and also there is a tremendous amount of increase in accuracy after using the FIS.

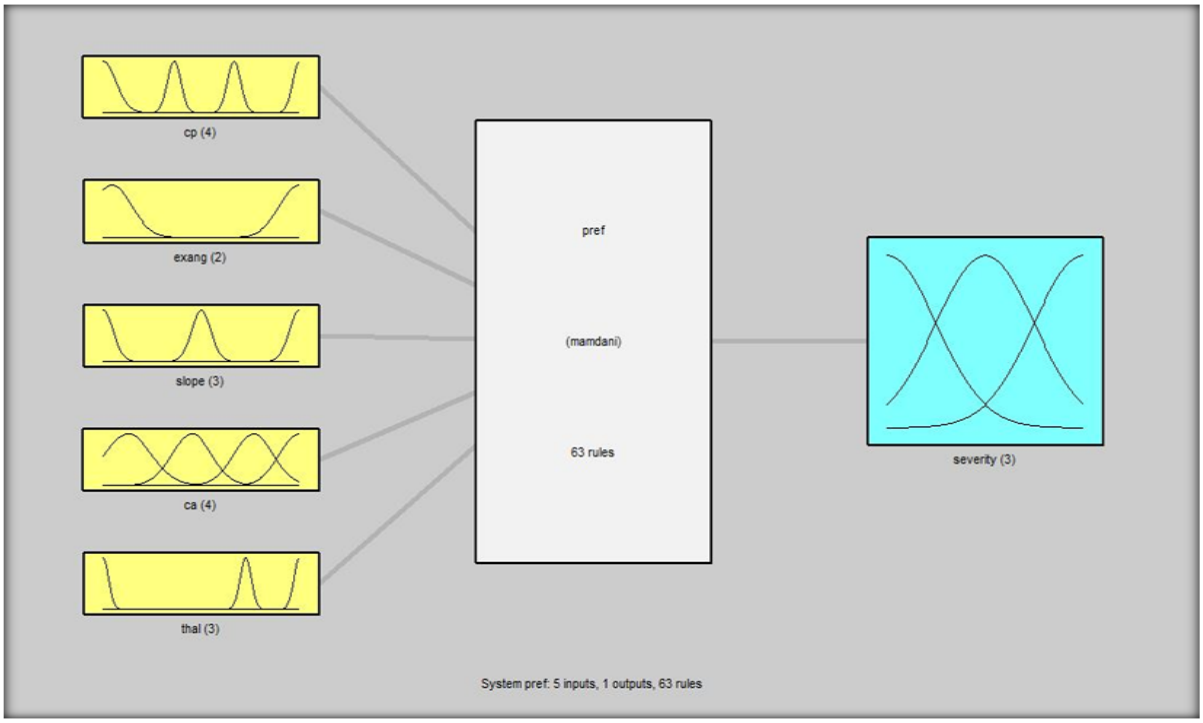


Figure 3: FIS designed for severity prediction

Figure 5: Rule viewer of the FIS in severity prediction

Table 4: Comparing of the output value

Original output	Manual calculation	System output
1	1.32	1.4
1	1.54	2
1	1.4	1.4
1	1.38	1.4
1	1.25	1.4
1	1.13	1.4
1	1.4	1.4
1	1.0	1.4
1	1.0	1.4
1	1.2	2
2	2.11	2
2	2.13	2
2	2.42	2
2	2.0	2
2	2.0	2
2	2.42	1.9
2	2.0	2.1
2	2.31	2
2	2.01	2
2	2.0	2
3	3.11	2.5
3	3.18	2.5
3	2.29	2.6
3	2.68	2.6
3	3.11	2.78
3	3.22	2.6
3	2.55	2.6
3	3.0	2.98
3	3.0	2.5
3	3.0	2.6

The Table 4 compares the original output with the result obtained from the system and mathematical calculation. It is observed that there is no much variation in the outputs when compared which justifies the use of the proposed FIS with Rule base for predicting the severity of the disease.

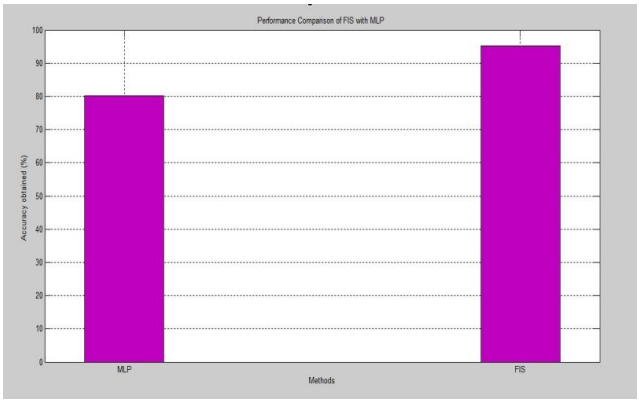


Figure 6 : Performance of FIS over MLP

The Figure 6 compares the performance of MLP over the proposed FIS with 63 rules framed. It is evident that the prediction accuracy has been increased from 80 to 90% which supports the use of FIS in predicting disease severity. The Figure 7 plot the training error with regard to number of epoch. It is observed that the training error has been reduced at the end of 50th epoch.

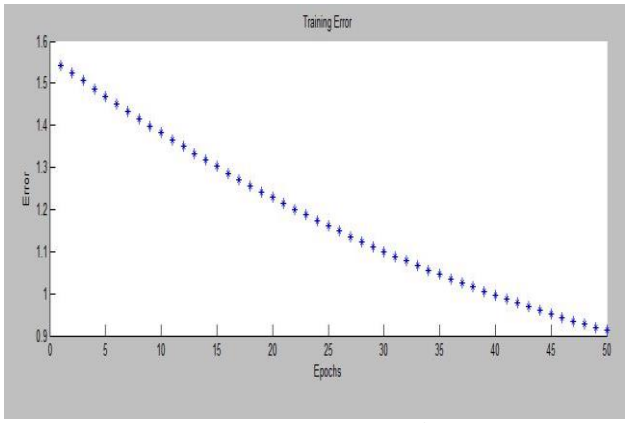


Figure 7: Training error Vs no. of epochs

7. Conclusion

The most important and difficult task in medicine is Medical diagnosis. The problem here is detecting a disease from several factors or symptoms, since it may lead to false assumptions with unpredictable results. The results obtained indicate that the proposed approach can be used to induce fuzzy rules from data by providing good balance between accuracy and readsability. Primary prevention is recommended for promoting healthy life style and habits through increased awareness and consciousness, to prevent development of any risk factors and a system to predict possibility of heart disease risk for prevention. The FIS designed from optimal data set has produced an accuracy of 95.23%. Hence it can be

suggested as a valid tool for medical practitioners in their preliminary stage.

References

[1] www.nytimes.com/2009/07/29/opinim/29hall.html.
[2] K.Rajeswari, "Prediction of Risk Score for Heart Disease in India using Machine Intelligence", *IPCSIT*, Vol 4, 2011
[3] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological and Life Sciences*, Vol 3(3), pp157-160,2007.
[4] Liangxiao. J, Harry.Z, Zhihua.C and Jiang.S "One Dependency Augmented Naïve Bayes", *ADMA*, pp 186-194, 2005
[5] Huan Liu and Hiroshi Motoda, Rudy Setiono and Zheng Zhao. "Feature Selection: An Everlasting Frontier in Data Mining", *JMLR: The 4th Workshop on Feature Selection and Data Mining*, 2010.
[6] Rafiah Awang and Palaniappan. S "Web based Heart Disease Decision Support System using Data Mining Classification Modeling techniques" , *Proceedings of iiWAS*, pp 177-187, 2007
[7] Carlos Ordonez, Edward Omincenski and Levien de Braal "Mining Constraint Association Rules to Predict Heart Disease", *Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society*, ISBN-0-7695-1119-8, 2001, pp: 433-440
[8] Deepika. N, "Association Rule for Classification of Heart Attack patients", *IJAEST*, Vol 11(2), pp 253-257, 2011.
[9] Durairaj.M, and Meena.K" A Hybrid Prediction System using Rough Sets and Artificial Neural Network", *International Journal of Innovative Technology and Creative Engineering*, Vol 1(7), July 2011.
[10] Durairaj.M, Sivagowry.S. "A Survey on Particle Swarm Optimization and Rough Set Theory in Feature Selection for Heart Disease Prediction." *International Journal of Computer Science and Mobile Computing (IJCSMC)* Vol 4 (3) (2015): 87-92.
[11] Sudha.A, Gayathri.p and Jaishankar. N "Utilization of Data Mining Approaches for prediction of life Threatening Disease Survivability", *IJAC* (0975-8887), Vol 14(17), March 2012.
[12] Rafiah Awang and Palaniappan. S "Intelligent Heart Disease Prediction System Using Data Mining techniques", *IJCSNS*, Vol 8(8), pp 343-350, Aug 2008
[13] Sivagowry.S, Dr. Durairaj. M. (2014). PSO-An Intellectual Technique for Feature Reduction on Heart Malady Anticipation Data. *International journal of Advanced Research in Computer Science and Software Engineering*, *4*(9), 610-621.
[14] Sivagowry.S, Dr. Durairaj.M. "A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction." *International Journal of Innovative Research in Computer and Communication Engineering* 2.11 (2014): 6457-6465.
[15] World Health Organization. Strategic priorities of the WHO Cardiovascular Disease programme. Available online at URL: http://www.who.int/whr/200. Last accessed February 2006.
[16] Chen A.H., "HDPS: Heart Disease Prediction System", *Computing in Cardiology*, ISSN 0276-6574, pp 557-560, 2011.
[17] en.wikipedia.org/wiki/myocardial_infarction
[18] Nidhi Bhatia and Kiran Jyothi, "A Novel Approach for heart disease diagnosis using Data Mining and Fuzzy logic", *IJCA*, Vol 54(17), pp 16-21, September 2012.
[19] Sivagowry, S., M. Durairaj, and A. Persia. "An empirical study on applying data mining techniques for the analysis and prediction of heart disease." *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*. IEEE, 2013.
[20] Durairaj.M, Sivagowry.S. "Feature Diminution by Using Particle Swarm Optimization for Envisaging the Heart Syndrome." *International Journal of Information Technology and Computer Science(IJITCS)* (7).2 (2015): 35-43.
[21] Muthukarupan S and Erib M.J, " A hybrid PSO based fuzzy expert system for the diagnosis of Coronary heart system", *Expert systems with Application*, Vol 39, pp 11657-11665, Elsevier, 2012
[22] Neera Pathania and Rithika, "FIS Rule based Heart Disease Prediction System to predict the risk level of Heart Disease", *IJCEM*, Vol 7, October 2014.
[23] Saeed Ayat, Asieh Khosravanian, "Identification and classification of Coronary Artery disease patients using Neuro-Fuzzy Inference system", *Journal of Mathematics and Computer Science*, Vol 13, pp 16-141, 2014.
[24] Saravana kumar, Tholkapia arasu, "Rough Set Theory and Fuzzy Logic based warehousing of Heterogeneous clinical database".
[25] AVS Kumar, " Generating rules for Advanced Fuzzy Resolution mechanism to diagnosis Heart Disease" , *International Journal of Computer Application*, Vol 77(11), pp 6-12, September 2012.
[26] Mohd. A.M. Abushariah, Assal A.M Alquadah, Omar Y.Adwan, Rana M.M. Yousef, " Automatic Heart Disease Diagnosis system based on Artificial Neural Network and ANFIS approach", *Journal of Software Engineering and Application*, pp 1055-1064, 2014
[27] UCI machine learning repository: http://archive.ics.uci.edu/ml/datasets/Heart+Disease: Last visited 8th August, 2014.
[28] Durairaj.M, Sivagowry.S, " An Intelligent Hybrid Quick Reduct Particle Swarm Optimization Algorithm for feature Reduction in Cardiac Disease Prediction", *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*, Volume 12(2), March-May 2015**,** pp.163-173

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Concept of a Work Management System in Nokia: Focusing on Goals Instead of Process Phases

Jari Lehto[1], Maarit Tihinen[2] and Päivi Parviainen[3]

[1] Nokia, Networks
PO BOX 1, 02022 Nokia Solutions and Networks, Finland
jari.lehto@nokia.com

[2] VTT Technical Research Centre of Finland Ltd
Oulu, PO BOX 1100, 90571, Finland
maarit.tihinen@vtt.fi

[3] VTT Technical Research Centre of Finland Ltd
Espoo, PO BOX 1000, 02044, Finland
paivi.parviainen@vtt.fi

Abstract

Complex systems development requires different ways of working than largely used static process oriented work. In practice, workers invent new ways of working to deal with appearing challenges. Thus, a company's processes and tools should support these new process paths. Instead of defining a specific flow to conduct the work, several options for work processes should be allowed. This paper introduced the main findings of a case study conducted in a complex product development environment. The goal of the case study was to improve the company's process support, based on the results of semi-structured interviews and the viewpoints of cognitive approach. The paper points out that it is important to focus on the goals for the work than the actual process phases and task descriptions. As a result of the study, the concept of a Work Management System (WMS) is proposed. The paper introduces this concept and discusses in details the main benefits of using a WMS.

Keywords: *Product process development, Work management system, Global product development, Cognitive approach.*

1. Introduction

Process-oriented approach for work management in organisations is not a new principle even if interesting to define and describe processes in organisations has been largely adopted in recent decades. In fact, focusing on processes to achieve quality and satisfy customers was promoted by Shewart in the 1930s [1, 2]. Furthermore, the globally adopted and well known quality management standard, the ISO 9001 Quality Management System [3], requires an organisation to apply a process approach to three types of actions: management, product and/or service realisation, and support. In addition, process owners shall be nominated with defined responsibilities and the authority to implement, maintain, and improve processes continuously. However, it has been pointed out [4] that organisations implementing some of the modern management concepts sometimes fail or do not achieve the expected results. In addition, frequent organisational changes are harmful, as they cause constant extra, and typically not resourced, updating and maintaining of the process management systems [5]. The same kinds of challenges and problems were encountered in our case study, where 40 semi-structured interviews were conducted at the case company premises in Finland. The main findings of the case study are introduced in this paper.

Based on [6] a process can be defined as follows: '*A process is a network of activities that are repeated in time, whose objective is to create value to external or internal customers.*' For example, the ISO 9001 standard can be seen to be based on this explication. Traditionally, organisations' product development processes are described and modelled by using process charts, where processes are introduced with models of static swimming courses. However, product development of complex systems differs from the currently used static process-oriented work. In practice, when developing complex systems workers invent new ways of working to deal with challenges that were not earlier defined (i.e., new process paths). Thus, it is difficult to specify a stable set of tasks or procedures for dynamic or unanticipated situations. Changes and unexpected events require creativity and human problem-solving skills in order to be overcome and solved [7]. In this paper, we emphasise the cognitive approaches to be taken into consideration while improving processes and developing process-oriented systems in organisations.

Cognitive work analysis (CWA) is an approach to work analysis focusing on how work can be done [8]. Instead of defining a specific flow to conduct the work, CWA emphasises that there are several options for work in terms of what to do, when to do it, and how to do it. While normative approaches focus on how work should be done, and descriptive approaches focus on how work is done, CWA focuses on the constraints that shape the work [8].

This paper introduces an industrial case study carried out at Nokia (URL: http://networks.nokia.com/), as well as a concept for a proposed solution — the work management system (WMS) — based on the case study. Nokia is a large global organization implementing both physical telecommunication network products and software for these products. The main products of Nokia are radio network elements. The case study focused on the development process used at Nokia, affecting the work of approximately 16,000 employees, though an estimated 2,000 people use the process actively in order to obtain instructions for their work. In this kind of environment, while improving a process, it is paramount to start with studies of actual work done. Thus, the ultimate goal of the Nokia case was to improve the working practices and process support by utilising the approaches of cognitive work analysis (CWA). This paper points out that it is important to focus on work goals instead of the actual process phases. In addition, the concept of a work management system is introduced and the main benefits of using a WMS are discussed in more detail. Even if the concept is still under pilot testing and being further developed, we argue that by focusing on the goals and decision points of the work the identified challenges can be managed.

This paper is structured in the following way. In the second chapter, background information relating to the case study, as well as relevant literature concerning the approaches of process management and modelling, business process systems, and cognitive methods are introduced. The third chapter presents the research design of our case study, in a descriptive way, with discussion of the main findings achieved. In the fourth chapter, the proposed solution, the concept of a work management system, is introduced and discussed. Finally, in the fifth chapter the main findings are summarized and concluded.

2. Background

The importance of adopting a process view and continuously improving processes has helped the process management philosophy to become a popular topic in the management literature in recent decades [4]. Based on the ISO 9001 standard, a process can be seen as a set of activities that are interrelated or that interact with one another. In practice, processes are interconnected, because the output from one process is often the input for another process, and resources are needed to transform inputs into outputs. The ISO process approach points out that a desired result is achieved more efficiently when activities and related resources are managed as a process. Process management is defined as the group of activities involved with organizing and monitoring the execution of a business process [9]. The term usually alludes to the management of business processes and manufacturing processes; even if business process management (BPM) and business process reengineering are interrelated but not identical [9]. Business process management is the discipline that combines knowledge from information technology and knowledge from management sciences and applies this to operational business processes [10]. Van der Aalst [10] identified four key BPM-related activities as follows: 1) model (creating a process model to be used for analysis or enactment); 2) enact (using a process model to control and support concrete cases); 3) analyse (analysing a process using a process model and/or event logs); and 4) manage (all other activities, e.g., adjusting the process, reallocating resources, or managing large collections of related process models).

In BPM the concept of a process model is fundamental [10]. Process models may be used to configure information systems, but they may also be used to analyse, understand, and improve the processes they describe. A process model aims to capture the different ways in which a case (i.e., process instance) can be handled. Process models assist in managing complexity by providing insight and by documenting procedures. Cross-organizational processes can only function properly if there is common agreement on the required interactions. The process-centric view in information systems has no consensus on notations and core capabilities. Despite the availability of established formal languages (e.g., Petri nets), industry has been pushing ad hoc or domain-specific languages. Furthermore, the control-flow perspective (modelling the ordering of activities) is often the backbone of a process model. Other perspectives, such as the resource perspective (modelling roles, organizational units, authorizations, etc.), the data perspective (modelling decisions, data, creation, forms, etc.), the time perspective (modelling durations, deadlines, etc.), and the function perspective (describing activities and related applications) are also essential for comprehensive process models. [10]

In the early Nineties Bernstein and Dayal [11] defined a repository as a '*shared database of information about engineered artefacts produced or used by an enterprise.*'

They described a repository manager as a database application that supports checkout/checkin, version and configuration management, notification, context management, and workflow control. Furthermore, Yan et al. [12] defined business process model repository software as software that supports the management of large collections of business process models. They also stated that software can assist in collections management by supporting common management functions such as storage, search, and version management of models. Thus, the software can also provide advanced functions that are specific for managing collections of process models such as managing the consistency of public and private processes. Accordingly, Dijkman et al. [13] pointed out that repository technology provides the actual infrastructure for storing a collection of process models. In addition, repositories are meant to support many management techniques (e.g., reuse, collection organization, query, search), such that they serve as the central point in an organization, from which the collection of business process models can be managed [13].

In software companies, there are typically several projects that differ greatly from each other, for example, in team size or expertise, product life cycle, or complexity. Therefore, the same process cannot apply to all types of projects. However, defining a new process for each project is not economically feasible. Thus, there is a need to define a series of processes as a process family, or to define a general process, including potential variation points, and then tailor it to each project. In the latter alternative, the tailoring costs are distributed among the projects [14]. Model-driven engineering (MDE) provides a formal framework for defining the models and transformations required for automated process tailoring. However, it requires formalization with various types of models specified and evolved, and it also requires support tools to be successful [15, 16]. The adoption of MDE approaches has been slow due to tool immaturity, as well as organisational and cultural issues [17].

Cognitive work analysis (CWA) is gaining ground in analysis, modelling, design, and evaluation of complex sociotechnical systems [8, 18]. Even though it is mostly used for interface design [8], it is a potential approach for identification of the properties of the work environment and of the workers. As Rasmussen et al. [19] stated, it is '*a methodological tool for planning field studies and data collection in various, actual work domains. It also serves as a means for a consistent analysis of collected empirical data and for representation of the results gained from empirical work studies.*' CWA assumes that in order to be able to design systems that work harmoniously with humans, one has to understand what the work actors do, their information behavior, the context in which they work, and the reasons for their actions [20]. During the case study, CWA was utilised for providing deeper understanding of the challenges identified in the current work context. In addition, the results were further analysed in order to elicit the main viewpoints and processing requirements for a renewal solution.

3. Research design

In this section, the case study and the research design is introduced. The starting point was that there is a need to renew the product development process in the case company. In practice, the development process in use contains various levels and phases, each focusing on different aspects of product development. The process involves all the stages of product development from ramp-up to ramp-down. In addition to this generic process, there are several process variants which have emerged due the individual requirements of the product programs that use the process. The process itself is large, containing 500 to 1,500 different process documents, depending on the program. For these reasons, tailoring of the process has experienced difficulties. However, the organizational culture in the case company provides much freedom for their employees to tailor the process to serve their needs in the best possible way.

In order to identify the unstructured problems, such as elements that work well in the current process and those that are experiencing challenges, in Autumn 2014 a total of 40 semi-structured interviews [21] were conducted involving various stakeholders from the case company. The semi-structured interview provides information in a more contextual and extensive way than the structured interview, and thus, it is a very useful interview technique in many software process development studies [22]. The interviewed stakeholders represent a wide range of roles (among others, project manager, program manager, product manager, product release manager, designer, tester, and architect) involved in product development in the case company. These interviews aimed towards identifying the requirements for the improved version of the development process and the portal (i.e., the repository) in which process-related information is stored. The interview data was analysed following the principles of thematic analysis, which is a method for identifying, analysing, and reporting themes (patterns) from data [23, 24]. In addition, interview findings were further analysed utilising the cognitive work analysis (CWA) method for understanding the root causes behind the main themes.

In this section, the research process and the main findings are introduced in detail. These findings formed a base for developing a concept of a work management system (WMS) in the case environment.

3.1 The main findings based on interviews

The data collected from the interviews revealed several themes that were formed utilising the thematic analysis method. The findings were divided into two different points of view. The first introduces findings from the process viewpoint: what were the most challenging topics experienced by respondents related to the following of current development processes (introduced in Section 3.1.1). The second introduces findings from the portal viewpoint: what kinds of challenges were identified by respondents relating to the current portal solution (introduced in Section 3.1.2). The interview findings are discussed in Section 3.1.3 and an overview of why cognitive work analysis was needed is provided.

3.1.1 Challenges in the current process descriptions

The top five themes related to the challenges experienced with the current product development process were as follows: 1) ambiguity; 2) heavy; 3) disconnection with stakeholders; 4) roles and responsibilities; and 5) unnecessary work. In this section, these topics are further discussed with examples extracted from the interviews in a manner that retains complete anonymity.

Ambiguity was seen as the most important theme with 19 individual respondents considering the current, generic version of the product development process. In practice, ambiguity was largely seen to stem from the fact that the current product development process aims towards providing guidance for all employees for conducting their work. Since the case company provides several products for the market, with unique features, differing constraints, and produced using various production methods and principles, it is, therefore, extremely challenging to provide detailed information for conducting tasks in the general version of the development process. This has resulted in product and program specific variants, which are often created by the people involved in the development of these separate products. The variants, however, are based on the general version of the development process.

The following citations from the content of the individual respondents illustrate ambiguity: '*It is unclear what elements of the process belong to general level processes and which to more detailed aspects'*, '*Large parts of the process are too abstract and hence not used*', '*Terminology can be ambiguous, which can be very challenging to inexperienced process users.'* In fact, ambiguity referred to process descriptions, terminology, acceptance criteria, etc. that were described in the process documentation either insufficiently or in such a manner that multiple interpretations of the information were possible. For example, there were several variants of a process used in different product lines and that's why terminology varied between projects, e.g., '*One has to rely on program specific variants for obtaining necessary information*'.

Heavy process was the second most important theme with 10 individual findings. A heavy process refers to the laborious nature of following the process. In addition, based on the findings, the process has become increasingly heavy during the years it has been developed in the case company. The following citations illustrate the findings: '*The process descriptions are not a good tool for gaining a deeper understanding of the process. They are too heavy for this purpose, which makes reading them a tiring task*'. '*Over the years the process has become increasingly heavy, since the mitigating actions have been added to the process when flaws have been found*'. However, it should also be noted that some of the findings related to this theme stemmed from individuals' negative attitudes towards all kinds of processes, for example, '*Following process details is laborious and a waste of time*'.

Disconnection between different stakeholders was the third most significant theme identified from the interviews with seven individual findings. The following findings illustrate this theme further: '*There is a disconnection between people who define the processes and those who use it. Hence, the process does not serve the actual needs and contains an exhaustive amount of additional responsibilities, such as meetings, that take up most of the work time*'. '*The link between the process and the work has been lost and those who are responsible for developing the process instructions might have been alienated a bit from the 'real' work*.'

In most situations, the disconnections were targeted between the people who define processes and those who use them. However, as described previously, several developers were willing to provide and use product and program specific variants of the process, while process people were trying to serve all employees with a general version of the development process. Furthermore, disconnection between the stakeholders performing different functions of the organization was also identified: '*HW and SW organizations are not working properly together. There is also disconnection between the testing organization and software architecture organization*'.

Roles and responsibilities related findings were indicated also in seven individual interviews. The following excerpts illustrate this theme further: '*The generic process explains the roles, but does not specify the actual persons to contact.*', '*Sometimes those labelled as process owners are not the 'real' process owners in a sense that they do not have time to put effort on process related matters.*' This theme refers to the ambiguities related to the responsibilities of different roles involved in the development work. In a large organisation, the links between person and role may diverge or blur, which results in uncertainty, gratuitous waiting, or work interruptions.

Unnecessary work was indicated to be a challenge in six individual interviews. Unnecessary work means work that does not contribute to actual product development, even if it consumes resources. Unnecessary work is generally referred to as waste in the literature discussing lean manufacturing [25] and lean software development [26, 27]. Thus, the unnecessary work not only increases expenses but also indicates the ineffectiveness of the processes. The following interview comments illustrate this theme further: '*Different tools are used at different sites resulting in unnecessary manual work.*', '*In some cases there is a lot of non-value producing work such as unnecessary documents that need to be prepared for a particular milestone.*' The first citation points out a real waste: unnecessary manual work. The second one can be a waste, but it can also be necessary work — perhaps important output — for other stakeholders. So, this type of process should make it clear, and in this way provide a sense of value for producing those kinds of documents.

3.1.2 Challenges relating to the current process portal solution

In this section, interview findings related to the current portal solution are introduced. In practice, interviewers had trouble differentiating the challenges of the process and the portal, because the current process has been implemented as a part of the portal. That was also the reason why there were fewer identified themes in the portal solution than in the process interviews. Only three main themes are introduced in detail.

The top three themes related to the challenges experienced with the current portal solution are as follows: 1) usability; 2) integrity; and 3) lack of visibility. In this section, these themes are discussed with a few examples extracted from the interviews.

Usability of the current portal solution was the most important theme that emerged from the interviews. This aspect was encountered in 29 interviews and covered several different sub-themes such as finding problems and search capabilities. In this case, usability refers not only to the aspects related to the use of the portal, but also to the ease of finding the actual portal from the web pages of the case company, or to the ease of finding information from the process portal. These sub-themes with illustrative comments are described as follows: '*Finding the link to the portal can be a challenge*', '*Checklists and templates are difficult to find and their location had to be asked from colleagues*'. Search capabilities and functionalities were mentioned as challenging: '*For example, the portal should utilize a task-based search system that clearly indicates what process to follow.*'

The second most important theme was ***integrity*** and related to forming a general view of information found in the current portal. This was indicated in eight individual interviews. This aspect is also related to the ambiguity of the process itself, such as ambiguity of the terminology, as discussed above. The comments below illustrate process portal-related integrity further: '*There should be a compact and clear introduction explaining what is found from the portal.*', '*One needs to know a lot about the process and where to find necessary information'*, '*It's missing how processes work together*'. Respondents expressed the need for visual representation of the processes or hovering with mouse-over instructions.

Lack of visibility pertains to information documented in the process, and information in the work products of those product programs or projects where people currently were working. This was seen as a challenge in three individual interviews. The following comments illustrate this aspect further: '*The portal should provide access to all individual units' internal process variants. This visibility should be applied also so that everyone that is working with the same aspects should be able to see their processes and ways of working.*', '*All the processes (e.g., delivery process and customer processes) should be found in the portal.*'

Other comments relating to the current portal solution covered the themes of re-use, simplicity, new ways of showing information, and process variants. The use of the portal as a tool for self-learning and re-use was seen as a challenge in two individual interviews. One respondent described this as follows: '*There should be practical examples on how certain things are documented or done. This could be re-used in future projects*'. Three other themes were identified in three separate interviews. An example of a comment relating to new ways of showing information was: '*Scrap the use of PowerPoint and build web pages that provide the content, for example, a program plan could be a web-based template updating*

online. All work is done online these days.' In practice, themes of simplicity and new ways of showing information can be also converted to a usability aspect.

3.1.3 Discussion of interview findings

The previously introduced main findings were very explicit and made the case for renewing the activities of the product development process in the case study environment. However, the specific findings relating to the challenges in the current process (3.1.1) and the challenges in the current portal solution (3.1.2) were greatly congruent, because the current process has been implemented as a part of the portal. In addition, the semi-structured interviews were not focused solely on the technical details such as features, possibilities, or constraints of the current portal solution.

In practice, the thematic analysis of the interviews provided the main themes and viewpoints to certain problems or challenging issues. However, in a large organisation there are several variants of a development process and different tools used across several sites. Thus, while interviewing people in different roles and from various sites, the actual root causes of the theme can be totally divergent. This truth does not reduce the significance of the findings instead it forces us to enlarge research activities and also the perspectives of the interview findings. Examples of these kinds of variety within the theme 'ambiguity' are as follows: '*One has to rely on program specific variants for obtaining necessary information.'*, '*Why are variants created? Could we just get rid of too many variants…'* Accordingly, comments such as '*hard to find correct or latest information, guidelines, and templates etc. or hard to identify the exact person to be contacted*' can be shown as themes of process ambiguity, heaviness, unnecessary work, or disconnection between different stakeholders, depending on the role and experience of the respondents. The use of the semi-structured interview method offers the means and possibilities to clarify each of the comments and sentences, but the actual root causes are often not clear for a respondent. For this reason, further investigations are needed.

Alongside the semi-structured interviews, it was recognised as important to be familiar with the general development process; process descriptions, phases, milestones, instructions, and templates etc. that are available from the process portal. This kind of process knowledge enables the researcher to ask accurate questions during the interview sessions. In addition, it was recognised that each respondent is a person working in their own way and feeling situations in their personal way. In practice, there are several options for conducting the work. Thus, the cognitive approach was taken into consideration while analysing the work activities of the respondents. A work analysis is a process of gathering information about work practices. Cognitive work analysis (CWA) provides a framework for work analysis that is seen as convenient for the analysis, design, and evaluation of complex sociotechnical systems [28, 29].

3.2 The cognitive approach for managing work

Generally, work analysis is seen as a systematic process of gathering information about work, tasks, and the relationships among tasks. On the most basic level, a workflow management system is any system that allows its users to setup, execute, monitor, and optimize different workflows. Nowadays, there are plenty of workflow management systems that serve different purposes, provide various features, and are based on different workflow languages. Accordingly, the literature shows various types of methods, ways, viewpoints, and principles for analysing and depicting work processes. Naikar et al. [8] proposed that a feasible level of task abstraction is used to bring in information that explains why and what for the task results are done and used. Thus, it is important to increase the understanding of the worker, instead of only showing the context-dependent practice. Hyysalo et al. [30] argued that while performing knowledge-intensive tasks and solving challenges creatively, a developer must understand both the current state and the goal state, and have a way to reach the goal. This understanding is the basis of problem solving and task implementation. Product development of complex systems differs from currently used static process-oriented work. Workers also invent new ways of working to deal with challenges not earlier defined [31]. Thus, it is difficult to specify a stable set of tasks or procedures for dynamic or unanticipated situations. For providing practical work support and promoting awareness of the progress, process should focus on information content [30].

The cognitive work analysis (CWA) method has been introduced in more detail as it provides an example framework of the cognitive approach for analysing work practices. CWA focuses on how the work can be done, and thus, it is gaining ground in analysis, modelling, design, and evaluation of complex sociotechnical systems [8]. Even though CWA is mostly used for interface design, it is a potential approach for identification of properties of the work environment and of the workers. In order to be able to design systems that work harmoniously with humans, one has to understand the work the actors do, their information behaviour, the context in which they work, and the reasons for their actions. CWA consists of five distinct phases [29]: 1) work domain analysis (WDA); 2) control

task analysis (ConTa); 3) strategies analysis (SA); 4) social organization and cooperation analysis (SOCA); and 5) worker competencies analysis (WCA).

In the literature, much of the research has focused on WDA, the first phase of CWA [32]. WDA focuses on the purposive and physical environments in which people operate. The purposive environment contains the reasons why the system exists and the physical environment includes the available resources. Together, these environments define the objectives that need to be achieved with the given resources. The aim of the second phase, ConTa, is to support people in dealing with recurring, known situations. ConTa focuses on what to do in the given work domain (purposive and physical environments) and complements WDA, since it identifies the activities that are necessary to achieve the goals with the given resources. The third phase, SA, focuses on identifying the different ways to accomplish the activity, i.e., it focuses on how the activity can be done. The fourth phase, SOCA, focuses on who is able to carry out the work requirements of the system, how it can be shared or distributed, and how it can be coordinated. And the fifth phase, WCA, focuses on the competencies that employees need to deal effectively with the work requirements of the system. In addition, there are several key concepts identified in each main analysis step of the CWA approach introduced previously. [8]

CWA provides a systematic framework for studying work content such as the main, required inputs/outputs, the necessary, critical, essential tasks/activities, the needed knowledge/skills/abilities, why the task results are necessary, and what they are used for, etc. Thus, CWA promotes understanding of the interaction between workers and information in the work context. In our case study, CWA was appropriately adapted during semi-structured interviews by collecting artefacts, eliciting current workflows, activities, tasks, tools used, environments, and asking respondents to represent or perform their work practices, for example. Utilising the cognitive approach for analysing the results of interview sessions increased the understanding of the worker and the work practices, instead of reporting themes without their context.

3.2.1 A renewed solution for supporting work practices

As previously introduced, a thematic analysis was utilised for eliciting the main themes regarding the challenges faced in the current supporting practices for the development processes in the case study. However, it was also recognised that the challenges might be represented in the same way even if their root causes were divergent. For this reason, the themes were further analysed using the CWA approach for understanding the root causes of each theme. Accordingly, the cognitive approach was taken into consideration, while further processing viewpoints and requirements for the proposed solution of a new work management system. In Table 1, example interpretations of utilising CWA in analysis and the provided viewpoints for the renewal solution are introduced.

The analysis of the interviews pointed out that there is a common, well-known and largely used decision-oriented product development process available from the case company portal. Although, there are several variants of the common process produced to meet the demands of specific product programs. The main purpose of decision-oriented development is to ensure that all promised goals are achieved. This means that there is much knowledge, information, and understanding behind each decision. Thus, the proposed solution should support a goal-oriented decision process by producing visibility for information, process work items, etc. The proposed solution should focus on supporting and advising the developer in their daily work, from feature screening, analysis, specification work, architecture design, software design, and testing work, keeping in mind the idea of a 24/7 work week. Thus, the proposed solution must support iterative and incremental work and has to offer the possibility for multiple users to work on the same task, goal, or other information online at the same time (real-time).

Table 1: Viewpoints for supporting work practices

Theme	*Example interpretations of utilizing the CWA approach*	*Viewpoints for a renewal solution (~requirements)*
Ambiguity	**Ambiguity** refers to terminology, process descriptions, work practices, acceptance criteria, etc. that are described in the process documentation either insufficiently, or in such a manner that multiple interpretations of the information are possible. In most cases, the root cause was derived from various product and program-specific variants. In addition, different programs used various tools and work practices. Thus, it was obvious that common process instructions were not familiar or were experienced as too abstract to use. The current solution covers all instructions; the common process instructions that are relevant for a few tasks, as well as detailed specific instructions utilized in other tasks or by persons on different teams.	The main viewpoint for avoiding ambiguity is to offer a single logical data repository that provides various role-based (or responsibility-based) views in a context-sensitive manner. The proposed solution should ensure that the latest information (not only guidelines and help, but also related technology and standards) is easily available for the right person. The purpose is to connect the tasks of individual developers with the proper guidance. In addition, the challenge refers to communication needs (e.g., chats, etc.) that should be easily available with help (hints, examples) or the provision of a society to facilitate communication. There should be support for easily finding the right person to get more information such as asking for advice or recommendations. The renewal solution should provide a view to one's own tasks in relation to the whole.
Heavy process	**Heavy process** refers to the laborious nature of following the process instructions and fulfilling the decision criteria. A heavy process also refers to the amount of information such as instructions, links, templates, figures, phases, steps, etc. Typically, the process includes a lot of history data and instructions from the several years during which it has been developed; terminology and site-specific instructions have been increasingly expanded during the years. In addition, this kind of common process is heavy and unfeasible to maintain; this involves contradictions in terms, too. The heavy process is one reason for misunderstandings, because the whole picture is unclear. The process can also be experienced as heavy if one utilizes only a small part of the whole	A heavy process can be avoided by reducing the amount of instructions and providing a feasible set of decision criteria. This can be done by focusing on the most relevant information and supporting the development work without separate instructions; some of the decision criteria can be included the work practices, for example. In addition, the guidance can be task specific without being laborious (e.g., YouTube videos). The proposed solution should provide easy and fast maintenance of the process; updates to instructions /templates have to be inherited to all needed places by utilizing links between the items. Furthermore, the solution should ensure traceability of information and history data should be easily available (e.g., by offering traceability items of each build and a baseline configuration).
Disconnection between different stakeholders **Roles and responsibilities**	The themes of **Disconnection between different stakeholders** and **Roles and responsibilities** refer to: 1) disconnection between people who define the processes and those who use them; 2) ambiguities related to the responsibilities of different roles involved in the work; and 3) disconnection between the work and decision criteria. For example, there were identified approver roles that suffered from being too far from the development work content. These kinds of roles and responsibilities have to be checked and adjusted in the current definitions to avoid extra work and misunderstandings	Avoiding these kinds of challenges can be done by developing the process towards a more lean and lightweight solution with process and criteria owners. Currently, part of the decision criterion is merely the 'Definition of Done' as an acceptance criterion. These kinds of criteria need to be integrated into actual work practices. After implementation of the described updates and lightening of the process, the identified challenges can be avoided by offering a single logical data repository that provides various role-based views in a context-sensitive manner (see Ambiguity).
Unnecessary work	**Unnecessary work** refers to: 1) difficulty in finding correct or the latest information; and 2) difficulties while different site-specific tools and practices are in use. The first indicates ineffectiveness and in this way increases expenses. The second can cause extra manual work with possibilities for mistakes if information is gathered from a tool and transferred to another tool or a document.	Unnecessary work can be avoided in the renewal solution if it ensures the latest information is easily available for the right person. The solution should support and advise daily work with the idea of a 24/7 work week. One of the main viewpoints for the renewal solution is that it should reduce manual work using automation. The automation can be provided with a single data repository, multiple views, and tools that are defined by models producing data to the work product.

4. The concept of a new work management system

In this section the concept of a proposed work management system (WMS) for the case company's environment is introduced.

The current process portal is used for obtaining necessary information about product development-related aspects of the process. This process portal is an online tool that can be found on the intranet: all process descriptions, procedures, instructions, and templates are available via the process portal. In practice, the process portal affects the work of all employees, either directly or indirectly. Thus, improvement and development actions that will be provided to the system are highly recommended, but are also very crucial. Thus, they have to be carefully planned.

The proposed solution of a WMS focuses on decision and decision-oriented process modelling with the purpose of making the decision visible. For example, the system explains what information is needed to make a decision. Decision points or milestones are regarded as basic building blocks of processes and processes themselves compose networks. Downstream decisions are divided into decision criteria, which are documented in such detail that the worker, in principle, can perform their tasks. Tasks are seen as information production activities to satisfy the decision making needed by the corresponding decision-makers.

As a solution, the model of process networks, processes, milestones and decision criteria extended with process communication and signalling structures, as well as roles, were defined. Based on the model, an application was generated, including the process networks and their components that can be instantiated as used process descriptions in the case company. Further, all the data included in the structures defined in the criteria can also be instantiated as project work products.

The solution includes a simple process model that describes only a process hierarchy and how processes communicate. A process is composed of decision points, which further are composed of decision criteria. The decision criteria include the actual guidance, explaining why and how the tasks shall be carried out. There are additional elements, such as common procedures and sub-processes, which are applied by other processes. Process elements are also communicating with each other by signalling and by data exchange.

As described in Figure 1, the WMS connects actual work and process guidance in a practical and seamless way. The data model of work products is defined based on requirements identified by stakeholders of the work domain. The model does not tell much about the sequence in which related information shall be brought in; the order is mostly defined by the process decision funnel.

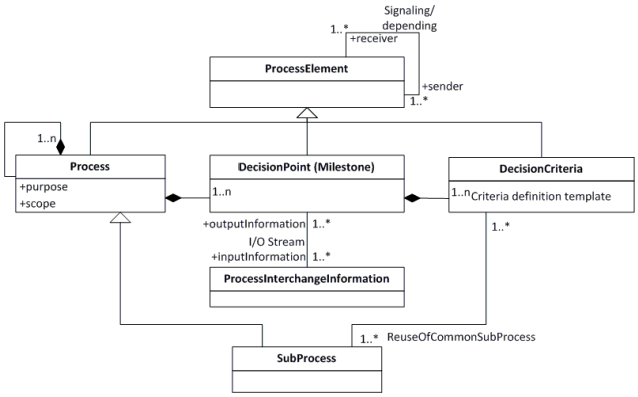


Fig. 1 High level model for a business process.

Furthermore, the order is multi-dimensional and cannot be defined as a clear sequence. The intermediary results are dependent and have effects on each other. This is typically a situation where the developer is the best one to determine the working order, and the order may change case by case. The decision funnel is described by decision points, which are further detailed by decision criteria. Typically, some issues are handled in several successive decision points by more detailed criterion, according to the principles of the funnel. From the work activity viewpoint, this means that from one topic some preliminary information is first collected, then more, and at the end the final set is defined. This explains how the criteria are task descriptions with a goal of produced information. The information is to be used when a decision is made. The produced information can also be used later, and thus, it flows down the process.

The WMS signals and forwards work products according to the decisions of other actors in the development community. The responsibilities of roles are an important part of process decomposition. Signalling synchronizes separate work activities or informs of special kinds of dependencies. Decision points may allow work products to be taken as source information by other parts of the process. The WMS may inform about conflict in work products under development and the developer may follow changes in the work product by expressing interest in following certain information (i.e., changes in the product).

Work management is a system for coordinating and recording the process of passing information, control

signals, and tasks from one worker or machine within a business to another. As technology advances, much work management has become partially automated and takes advantage of special software to make the process smoother. Thus, the WMS improves work performance significantly.

5. Conclusions

This paper introduced a case study that was conducted during 2014–2015 in the Nokia environment. The main goal of the case study was to analyse challenges in the current work context, including process support, in order to provide an improved solution that would effectively support the work practices in the various roles of the company. In addition, the proposed solution, the concept of a work management system (WMS), was introduced and discussed in detail.

The purpose of the WMS is to connect the tasks of individual developers with proper guidance, inform other activities, follow-up, and to gather data for lessons learned. The paper pointed out and discussed several benefits that are enabled with the proposed WMS concept. The proposed WMS supports different ways of performing development work to achieve goals. As stated in the literature [31, 33], decision-oriented, transparent processes can raise a developer's awareness of the people working on a particular decision.

As introduced in the paper, this kind of goal-oriented guidance also promotes innovative work practices. The WMS also supports the main features of decision-oriented development with embedded decision criteria and guidance for a developer and in this way provides visibility to the information needed for making a specific decision. Accordingly, the WMS provides visibility to the progress of tasks and work via various views of traced work products or items. Keeping track of decisions, rationale, and the effects of decisions on software products are also advised in the literature [31, 34]. Traceability and change control functionalities of the WMS enable controlled change decisions with visibility to the rationale behind the changes. In addition, this kind of context-sensitive role of the WMS reduces idle time and unnecessary work; simply expressed, reduces waste in an organisation. Furthermore, a static structure at the highest level of the process hierarchy reduces maintenance cost. The updates and modifications are done once to a particular location of a single logical data repository.

Although in this paper a WMS is introduced mainly as a conceptual framework, a proof of concept (PoC) has been developed. This PoC has been demonstrated while validating the benefits of the proposed WMS. At the moment, the concept is under further development in the case organisation. The results introduced in this paper should interest academics and practitioners as they indicate how the cognitive approach can be utilised for studying current practices in a large global company. The study also provides valuable insights for academics, as it combines different approaches, such as theories from the information sciences, technology, process and business oriented viewpoints, and cognitive approaches. For practitioners, the introduced case study with the proposed WMS provides a better understanding of the context of the work by defining the real needs of stakeholders, processes, activities, and tasks. In this paper the importance of focusing on work goals, instead of on the actual process phases is introduced and discussed with the concept of a work management system. The benefits of a WMS are also concluded and discussed.

Acknowledgments

This case study was conducted during the PROMES ITEA2 project, number 11013 (PROMES, Processes Models for Engineering of Embedded Systems). The authors would like to thank the support of ITEA (ITEA, Information Technology for European Advancement) and Tekes (The Finnish Funding Agency for Innovation) for enabling the research.

References

[1] W. A. Shewhart, Statistical Method from the Viewpoint of Quality Control, Washington, Graduate School of Agriculture, 1939.
[2] W. A. Shewhart, Economic Control of Quality of Manufactured Product, ASQ Quality Press, 1931.
[3] ISO 9001:2015 Standard, "Quality management systems – requirements", September 2015.
[4] P. Cronemyr and M. Danielsson, "Process management 1-2-3–a maturity model and diagnostics tool", Total Quality Management & Business Excellence, vol. 24, (7-8), 2013, pp. 933-944.
[5] J. J. Dahlgaard, J. Pettersen and S. M. Dahlgaard-Park, "Quality and lean health care: A system for assessing and improving the health of healthcare organisations", Total Quality Management & Business Excellence, vol. 22, (6), 2011, pp. 673-689.
[6] B. Bergman and B. Klefsjö, Quality from Customer Needs to Customer Satisfaction, Studentlitteratur, 2010.
[7] A. Abecker, S. Dioudis, L. van Elst, C. Houy, M. Legal, G. Mentzas, S. Müller and G. Papavassiliou, "Enabling workflow-embedded OM access with the DECOR toolkit", Knowledge Management and Organizational Memories, Springer, 2002, pp. 63-74.

[8] N. Naikar, R. Hopcroft and A. Moylan, Work Domain Analysis: Theoretical Concepts and Methodology, 2005.
[9] J. Becker, M. Kugeler and M. Rosemann, Process Management: A Guide for the Design of Business Processes, Springer Science & Business Media, 2013.
[10] van der Aalst, Wil MP, "Business process management: A comprehensive survey", ISRN Software Engineering, 2013.
[11] P. A. Bernstein and U. Dayal, "An overview of repository technology", In Proceedings of VLDB, vol. 94, 1994, pp. 705-713.
[12] Z. Yan, R. Dijkman and P. Grefen, "Business process model repositories–Framework and survey", Information and Software Technology, vol. 54, (4), 2012, pp. 380-395.
[13] R. M. Dijkman, M. La Rosa and H. A. Reijers, "Managing large collections of business process models-current techniques and challenges", Computers in Industry, vol. 63, (2), 2012, pp. 91-97.
[14] P. Kruchten, The Rational Unified Process: An Introduction, Addison-Wesley Professional, 2004.
[15] M. Bastarrica, J. Simmonds and L. Silvestre, A Megamodel for Process Tailoring and Evolution, 2014.
[16] J. Simmonds, D. Perovich, M. C. Bastarrica and L. Silvestre, "A megamodel for software process line modeling and evolution", Proceedings of MODELS, 2015.
[17] J. Whittle, J. Hutchinson, M. Rouncefield, H. Burden and R. Heldal, "Industrial adoption of model-driven engineering: Are the tools really the problem?", In Proceedings of the Model-Driven Engineering Languages and Systems, 2013, pp. 1-17.
[18] P. M. Sanderson, Cognitive Work Analysis, 2003.
[19] J. Rasmussen, A. M. Pejtersen and K. Schmidt, Taxonomy for Cognitive Work Analysis, 1990.
[20] R. Fidel, A. Mark Pejtersen, B. Cleal and H. Bruce, "A multidimensional approach to the study of human- information interaction: A case study of collaborative information retrieval", Journal of the American Society for Information Science and Technology, vol. 55, (11), 2004, pp. 939-953.
[21] P. N. Ghauri and K. Grønhaug, Research Methods in Business Studies: A Practical Guide, Pearson Education, 2005.
[22] M. Vierimaa, J. Ronkainen, O. Salo, T. Sandelin, M. Tihinen, B. Freimut and P. Parviainen, "Comprehensive collection and utilisation of software measurement data", VTT Publications, 2001.
[23] V. Braun and V. Clarke, "Using thematic analysis in psychology", Qualitative Research in Psychology, vol. 3, (2), 2006, pp. 77-101.
[24] M. Vaismoradi, H. Turunen and T. Bondas, "Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study", Nurs. Health Sci., vol. 15, (3), 2013, pp. 398-405.
[25] M. Poppendieck and T. Poppendieck, Implementing Lean Software Development: From Concept to Cash, Pearson Education, 2007.
[26] J. P. Womack and D. T. Jones, "Beyond toyota: How to root out waste and pursue perfection", Harvard Business Review, vol. 74, (5), 1996, pp. 140-.
[27] J. P. Womack and D. T. Jones, Lean Thinking: Banish Waste and Create Wealth in Your Corporation, Simon and Schuster, 2010.
[28] J. Rasmussen, A. M. Pejtersen and L. P. Goodstein, Cognitive Systems Engineering, Wiley, 1994.
[29] K. J. Vicente, Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work, CRC Press, 1999.
[30] J. Hyysalo, J. Lehto, S. Aaramaa and M. Kelanti, "Supporting cognitive work in software development workflows", In Proceedings of the Product-Focused Software Process Improvement PROFES, Springer, 2013, pp. 20-34.
[31] J. Hyysalo, M. Kelanti, J. Lehto, P. Kuvaja and M. Oivo, "Software development as a decision-oriented process", 2014, pp. 132-147.
[32] N. Naikar, "An examination of the key concepts of the five phases of cognitive work analysis with examples from a familiar system", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2006, pp. 447-451.
[33] A. Aurum and C. Wohlin, "The fundamental nature of requirements engineering activities as a decision-making process", Information and Software Technology, vol. 45, (14), 2003, pp. 945-954.
[34] G. Ruhe, "Software engineering decision support – a new paradigm for learning software organizations", Advances in Learning Software Organization, Lecture Notes in Computer Science, Springer-Verlag, 2003, pp. 104-113.

Lic.Sc. Jari A Lehto has extensive managerial experience from industrial enterprises and currently he is working in a large telecommunication company in method development. He is responsible for supporting business units in the area of architecture and system design especially in requirements engineering. He has contributed to research of multi-site collaboration practices. He has graduated in information processing science 1993 (Ph. Lic.) in University of Oulu, Finland.

Dr. Maarit Tihinen is a Senior Scientist in VTT Technical Research Centre of Finland. She graduated in the department of mathematics from the University of Oulu in 1991. After that she completed her secondary subject thesis in 2001, and received her PhD in 2014 in information processing science from the University of Oulu, Finland. Tihinen has worked in several national and international research and customer projects and has written scientific publications for international software engineering conferences and journals. Her research interest includes quality and process improvement, measurements and metrics, Industrial Internet and digital service development.

Dr. Päivi Parviainen received her M.Sc. in information processing science from the University of Oulu in 1996 and her PhD in information processing science from the University of Oulu in 2013. She is currently working as a Principal Scientist in the digital services in context team at VTT in Espoo, Finland. She has worked at VTT since 1995. Over the years, she has authored over 30 publications. Her research interests include digitalisation, product and digital service development practices.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A mission location recommender system to missioner by using clustering based collaborative filtering

Razieh Qiasi[1], Seyyed Hassan Hani-Zavarei[2], Behrooz Minaei-Bidgoli[3]
[1]Department of Computer, Sama technical and vocational training collage, Islamic Azad University, Qom Branch, Qom, Iran
raziehghiasi@gmail.com
[2]Department of Information Technology, University of Qom
Qom, Iran
Hasanhani@yahoo.com
[3]Department of Computer Engineering, University of Science and Technology
Qom, Iran
minaeibi@cse.mcu.ed

Abstract

By expansion of religion mission boards to further parts of Iran, and also many different mission needs and increasing number of missions and mission locations, traditional and manual methods of missioner dispatch are not fast and accurate enough for dispatching manager's needs anymore. So, there is a need for an intelligent system which can improve dispatching programs by assisting the missioners in selecting the suitable location. Application of recommender systems is a suitable solution to this problem. Collaborative filtering is the most commonly used and effective recommendation technique among different types of recommender systems.

This paper presents a mission location recommender system based on collaborative filtering method. Traditional CF method is not scalable for the increasing number of missioners. To address this issue, this paper proposes developing a mission location recommender system based on clustering techniques followed by collaborative filtering. The experimental results show that the cluster based collaborative filtering has acceptable performance and it is the most accurate and scalable user based CF.

Keywords: Dispatching missioner, Recommender system, Collaborative filtering, clustering.

1. INTRODUCTION

Religion propaganda has an effective role in Iranians cultural and Islamic knowledge improvements and also it is important in exploitation of religion sciences scholars for the purpose of publication and spreading of religious learning. Because of religion propaganda's importance, there are several centers responsible of dispatching scholars and clerics to missions in Iran.
Suitable mission locations' selection is affected by several parameters, such as spreading of religion mission board in further parts of Iran, many different missions' needs and increasing number of missioners and mission locations, but traditional and manual methods of mission dispatch are not fast and accurate enough for dispatching managers' needs. Therefore there is a need for an intelligent system which can assist the missioners in selecting the suitable mission location and so improve dispatching programs. Recommender systems are a suitable solution for this problem.
Recommender systems are decision supports which present information about items with respect to user preferences by analyzing users' prior behavior. Generally three kinds of methods are applied in recommender systems: content filtering, collaborative filtering and hybrid method. Collaborative filtering is the most popular one among these methods. A collaborative filtering system's basic idea is to generate recommendations based on similar past users' experiences [20]. Collaborative filtering method can be memory based or model based.
Memory based collaborative filtering make recommendations on all of the gathered data. In this method the newly generated data can also be taken into account for recommendation, so its recommendations can be highly accurate. In model based collaborative filtering, first a model on all of the offline data must be constructed and then this model will be loaded to the memory to generate online recommendation results. While making some concessions on accuracy, this method significantly improves system's scalability.
Memory based techniques are quite successful in real world applications, because they are easy to understand, implement and work well in many real world situations. However, there are some problems that limit the application of memory based techniques like user-item rating matrix which will result in a scarcity matrix, especially in large scale applications, that each user only rates a small set of a large database items.
To overcome the weaknesses of memory-based techniques, researchers has focused on hybrid memory-based and model-based approaches with the aim of seeking more accurate, yet more efficient methods [3,5,12,15].

This paper proposes a hybrid memory and model based approach for building a mission location recommender system. This approach uses clustering techniques to identify the communities of similar missioners based on their rating data and uses these communities as a mechanism to make the recommendations.
Our efforts in this paper are aimed towards applying existing recommendation methods in propaganda domain, which is a new domain of issues. So, we are not going to propose a new method in recommendation systems.
The rest of this paper is consisted of the following sections. Section 2 summarizes the related works. The research methodology used in this study is reported in sections 3. Evaluation metrics used in this study are discussed in section4. Section 5 presents empirical results. Finally, Section 6 summarizes our conclusions of this work and suggests future research directions.

2. RELATED WORK

In case of having user clusters, traditional CF algorithms can be operated on the clusters instead of the whole user-item matrix. By reducing the dimensions of user-item rating matrix and therefore avoiding the data sparsely problem, this approach can provide better recommendation results in terms of accuracy and can improve the online performance of CF algorithms. So far, many researchers have used clustering to improve the scalability and sparsely problem. In the following, we'll describe more some of these researches.

A. Kohrs et al. [1] presents a novel algorithm for collaborative filtering based on hierarchical clustering which tries to balance robustness and accuracy of predictions and experimentally show that it is especially efficient in dealing with *bootstrapping* and *new user* situations.

S.H.S. Chee et al. [18] developed an efficient collaborative filtering method called RecTree that applies clustering techniques which create cohesive cliques economically. RecTree achieves better scale-up in comparison to other memory based collaborative filters by seeking advisors only within a clique rather than the entire database.

Sarwar et al. [3] presented a clustering-based algorithm that is suitable for a large data set.

Bridge et al. [4] generalized an existing clustering technique and applied it to a collaborative recommender's dataset to reduce cardinality and sparsely. They systematically tested several variations, by exploring the value of partitioning and grouping the data.

In Kelleher et al. [11] a collaborative recommender is presented that uses a user-based model to predict user ratings for specified items. The model comprises summary rating information derived from a hierarchical clustering of the users. They compare their algorithm with several others and show that its accuracy is good and its coverage is maximal. They also showed that the proposed algorithm is very efficient. prediction time in this method grows independently of the number of ratings and items and only grows logarithmically with respect to the number of users.

Xue et al. [8] presented a novel approach that combines the advantages of memory based collaborative filtering and model based collaborative filtering approaches by introducing a smoothing-based method. In this approach, clusters generated from the training data provide the basis for data smoothing and neighborhood selection. As a result, they provide higher accuracy as well as increased efficiency in recommendations. Their empirical studies on EachMovie and MovieLens datasets shows that the proposed approach consistently outperforms other user based traditional collaborative filtering algorithms.

Rashid, A.M. et al. [2] proposed ClustKnn, a simple and intuitive algorithm that is well suited for large data sets. First, by building a straightforward but efficient clustering model, this method tremendously compresses data. Recommendations are then generated quickly by using a simple Nearest Neighbor-based approach. The feasibility of ClustKnn has been demonstrated both analytically and empirically. They have done a comparison with a number of other popular CF algorithms which shows that apart from being highly scalable and intuitive ClustKnn provides very good recommender accuracy as well.

Mittal et al. [14] proposed a framework based on data partitioning/clustering algorithm application on ratings dataset followed by collaborative filtering for developing a movie recommender system. This system reduces the computation time considerably and increases prediction accuracy.

3. METHODOLOGY

This section provides details of the purposed method for constructing mission location recommender system. This method has two phases: offline phase (user clustering) and online phase (generation of prediction and recommendation).

3.1 User Clustering

User clustering techniques work by identifying groups of users who appear to have similar ratings (see Fig.1). Once the clusters are created, predictions for a target user can be made by averaging the opinions of the other users in that cluster. There are many algorithms that can be used to create clusters. In this paper, a *TwoStep* algorithm is selected.

TwoStep Clustering is a two-step clustering method. The first step compresses the raw input data into a manageable set of sub-clusters by making a single pass through the data. The second step uses a hierarchical

clustering method to progressively merge the sub-clusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering does not require the number of to-be-selected-clusters ahead of time.

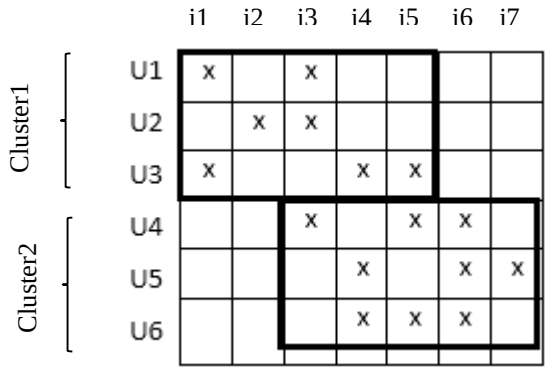


Fig. 1: User Clustering

This method uses a log -likelihood distance measure, to accommodate both symbolic and range fields. It is a probability-based distance. The distance between two clusters is related to the decrease in log-likelihood as they are combined into one cluster. We have assumed log-likelihood normal distributions for range fields and multinomial distributions for symbolic fields in our calculations. It is also assumed that the fields are independent of each other and so are the records. The distance between clusters i and j is defined as:

$$d(i,j) = \xi + \xi_j - \xi_{\langle i,j \rangle} \quad (1)$$

Where

$$\xi_u = -N_u\left(\sum_{k=1}^{k^A} \frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{uk}^2) + \sum_{k=1}^{k^B} \hat{E}_{uk}\right) \quad (2)$$

and

$$\hat{E}_{uk} = -\sum_{l=1}^{L_k} \frac{N_{ukl}}{N_u} log \frac{N_{ukl}}{N_u} \quad (3)$$

In the above equations, k^A is the number of range type input fields, k^B is the number of symbolic type input fields, L_k is the number of categories for the kth symbolic field, $N_u$is the number of records in cluster u, N_{ukl} is the number of records in cluster u which belongs to the lth category of the kth symbolic field, $\hat{\sigma}_k^2$ is the estimated variance of the kth continuous variable for all records, $\hat{\sigma}_{uk}^2$ is the estimated variance of the kth continuous variable for records in the vth cluster, and $\langle i, j \rangle$ is an index representing the cluster formed by combining clusters i and j.

If we ignore $\hat{\sigma}_k^2$ in the equations for ξ_u, the distance between clusters i and j would be exactly the decrease in log-likelihood when the two clusters are combined. The $\hat{\sigma}_k^2$ term is added to solve the problem caused by $\hat{\sigma}_{uk}^2 = 0$, which results in natural logarithm being undefined. (This would occur, for example, when a cluster has only one case)[19].

3.2 Generation of Prediction and Recommendation

The main task of rating prediction is finding nearest neighbors for active users. Finding the nearest neighbors requires computing the similarity between users. Therefore, this section includes these three main steps: similarity computation, neighborhood selection and the processes involved in rating prediction.

3.2.1 Similarity Computation

The notion of similarity is used to identify users that have common "preferences". As mentioned above, traditional memory based collaborative filtering searches the whole ratings database to find the most similar users. Whereas in the method used in this paper, similarity of active user is computed by the members of the cluster which it blongs to. Therefore, execution time is reduced and scalability and sparsely problems are resolved too.

There are several methods to measure similarity among which Pearson's correlation and cosine vector similarity are widely used in collaborative filtering [6,7].

- Pearson correlation coefficient (PC): This metric measures the degree of association between ratings' patterns using a value between -1 and +1. A positive value is the evidence of a general trend where high ratings of user U are associated with high ratings of V and low ratings of U tend to be associated with low ratings of V (a negative value for the correlation implies the inverse of this association). PC can be computed by:

$$PC(u,v) = \frac{\sum_i (r_{ui} - \bar{r_i})(r_{vi} - \bar{r_i})}{\sqrt{\sum_i (r_{ui} - \bar{r_i})^2}\sqrt{\sum_i (r_{vi} - \bar{r_v})^2}} \quad (4)$$

Here $r_{u,i}$ $(r_{v,i})$ denotes the rating of user u (v) on item i, $\bar{r_i}$ is the average rating of the i-th item.

- Cosine measure: This metric defines the similarity between two users as the cosine of the angle between the rating vectors, with values between 0 and 1. A larger value means a higher similarity for the ratings (the two vectors are closer). The cosine similarity of users U and V is defined as:

$$Cosine\ (u,v) = \frac{\sum_i r_{ui} r_{vi}}{\sqrt{\sum_i {r_{ui}}^2}\sqrt{\sum_i {r_{vi}}^2}} \quad (5)$$

3.2.2 Neighborhood Selection

The next stage is selection of the neighbors who will serve as recommenders. Here, the entire cluster that user belongs to, can be selected as user's neighborhood. For a fair comparison we have recorded the number of neighbors used for prediction computation for each user and forced our basic CF algorithm to use same number of neighbors for prediction generation. Selection of the neighbors is normally done in two steps [9,17]:

- Threshold-based selection, in which the users whose similarity exceeds a certain threshold value are considered as neighbors of the target user.
- The top-n technique, in which n-best neighbors are selected; n's value is given ahead.

3.2.3 Prediction Rating

When a subset of the nearest neighbors of the active user are selected, predictions are generated based on a weighted aggregate of their ratings. Most used aggregating functions are weighted sum and simple weighted average. To make the prediction for the active user *u* on an item *i*, weighted sum is computed using all the ratings of the neighbors on that item by the following formula:

$$Pr_{u,i} = \overline{r_u} + \frac{\sum_{k \epsilon K} Sim(u,k)(r_{ki} - \overline{r_k})}{\sum_{k \epsilon K} |sim(u,k)|} \quad (6)$$

4. EVALUATION

Evaluation is one of the key aspects in recommender systems. There has been considerable research in the area of recommender system's evaluation focused on accuracy and performance [6,10]. These researches introduce several metrics for assessing the accuracy of collaborative filtering methods [13]. These metrics are divided into two main categories: statistical accuracy metrics and decision-support accuracy metrics.

Statistical accuracy metrics: Statistical accuracy metrics evaluate the accuracy of a prediction algorithm by comparing the numerical deviation of the predicted ratings from the respective actual user ratings. Some of the frequently used metrics are mean absolute error (MAE), root mean squared error (RMSE).

Mean absolute error (MAE) is a quantity used to measure how close predicted ratings are to the actual rating as shown in (Eq.7). Root mean squared error (RMSE) amplifies the contributions of the absolute errors between the predictions and the true values as shown in Eq. (8).

$$MAE = \frac{\sum_{u,i} |Pr_{u,i} - r_{u,i}|}{U * I} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{u.i} (Pr_{u,i} - r_{u,i})^2}{U * I}} \quad (8)$$

Where number of users, number of items, predicted rating and true rating are represented by U, I, $Pr_{u,i}$ & $r_{u,i}$. As lower as the MAE or RMSE becomes, the more accurate the predictions would be, thus formulating better recommendations will be possible.

Decision-support accuracy metrics: Decision-support accuracy metrics evaluate how effectively predictions help a user to select high-quality items. Some of the frequently used metrics are recall, precision and F-Measure.

The Precision metric measures the share of successful recommendations from the total number of computed recommendations (Eq. 9), while the Recall metric is the ratio of the number of ratings correctly predicted over the total test data (Eq.10).

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

Here, TP is an interesting item that is recommended to the user, FN is an interesting item that is not recommended to the user and FP is an uninteresting item that is recommended to the user.

There are some drawbacks with using only two metrics of recall and precision. For example, with increasing the size of recommendation list, recall will increment, while precision will decrement. The F-measure has been used to alleviate these problems through applying the harmonic average of precision and recall which is defined as follows:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

5. EXPERIMENTAL RESULT

5.1 Dataset

The data for this study is drawn from a dataset of Islamic center in Iran. In order to construct missioner rate matrix, we used the overall dispatch of a missioner to a specific location as the rate of missioner for that location. All ratings are integer values between one as the lowest value and twenty two as the highest value. Also, this data includes all missioners who dispatch at least one location. Therefore, 11850 ratings is accessible for 31 locations (31 state of Iran) and 8030 missioners. The sparsely (percentage of zero values in the missioner-location matrix) is 95.24%.

5.2 Experimental Procedure

The first stage in this study is creation of user clusters. But before that, in order to improve the quality of clustering, missioner-location matrix is normalized by using Gaussian normalization. In this method, the normalized rating for item *i* by user u,$\hat{r}_{u,i}$is computed as follows [16]:

$$\hat{r}_{u,i} = \frac{r_{u,i} - \overline{r_u}}{\sqrt{\sum_i (r_{u,i} - \overline{r_u})^2}} \quad (12)$$

Where $r_{u,i}$ stands for the rating of item i by user u, and $\overline{r_u}$ stands for the average rating for user u.

Then missioners have been partitioned into 18 clusters by using *TwoStep* method and based on missioner rate data. Figure2 shows distribution of cluster proportion for *TwoStep* Clustering.

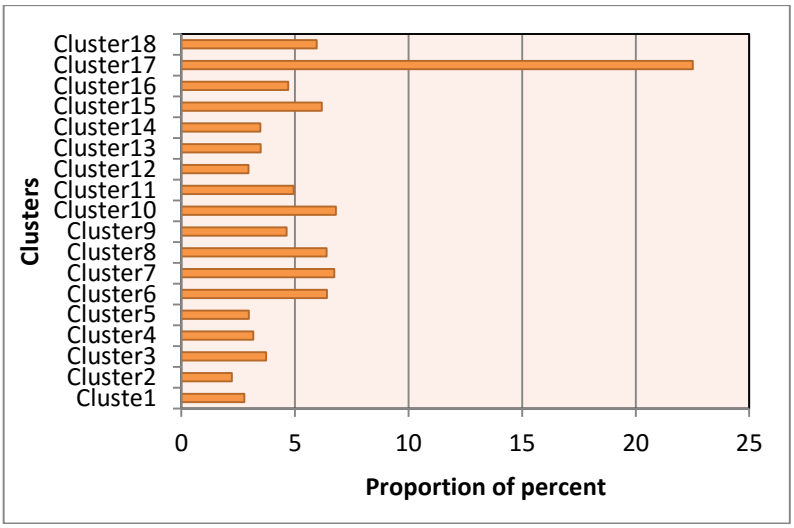


Fig.2: Distribution of cluster proportion for *TwoStep* Clustering.

As Fig. 2 shows, the *TwoStep* model tends to keep the size of different clusters balanced, which will create a better interpretation, capturing wider variations in the missioner's behavior.

In order to simplify the interpretation of the clusters, distribution of mission location in clusters' diagram is also shown in figure 3. As Fig. 3 shows, each cluster includes a location that has been requested and dispatched more than other locations. For example, in cluster1, many missioners dispatched to Gilan.

In the next phase, to conduct prediction and create recommendation, firstly the data set splits into two training and testing sets. We have chosen randomly 1606 missioners (20%) as the test set and the rest of the missioners as the training set (80%). Finally, with respect to the presented methodology in section 3, the recommender system is generated and the suitable mission location is recommended to the missioner.

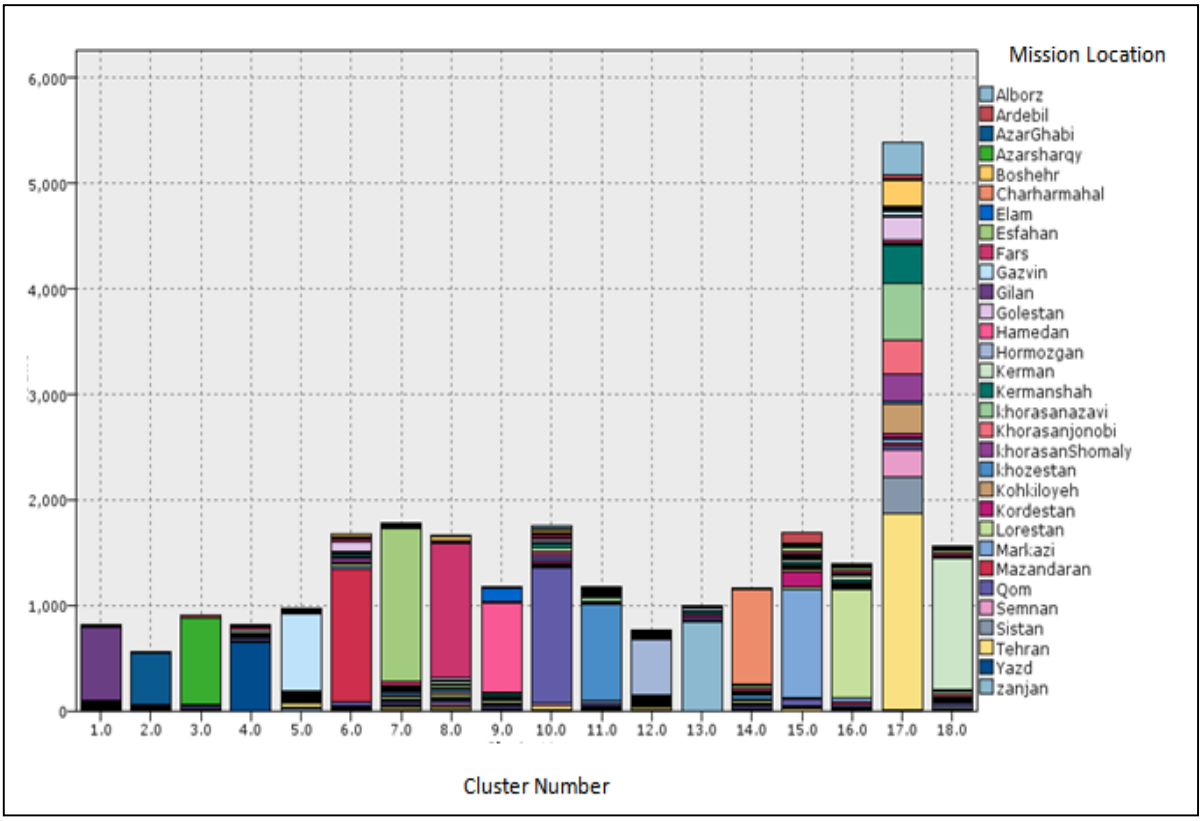
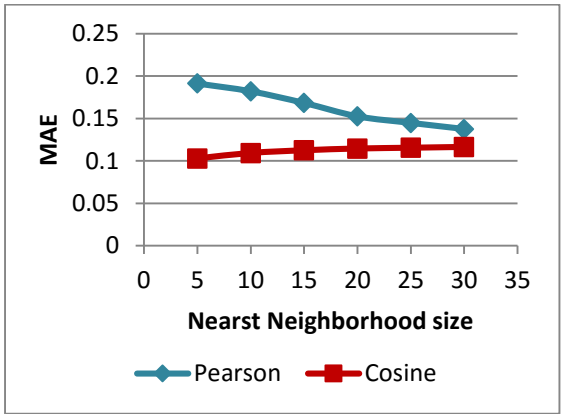


Fig. 3: Distribution of mission location in clusters.

5.3 Results Analysis

The size of neighborhood can have a significant impact on prediction quality; we built up our experiment by varying the neighborhood size from 5 to 30 and validating the predictions' efficiency by computing the MAE and RMSE metrics. Figure4 illustrates the sensitivity of the algorithms in relation to the different numbers of neighbors.



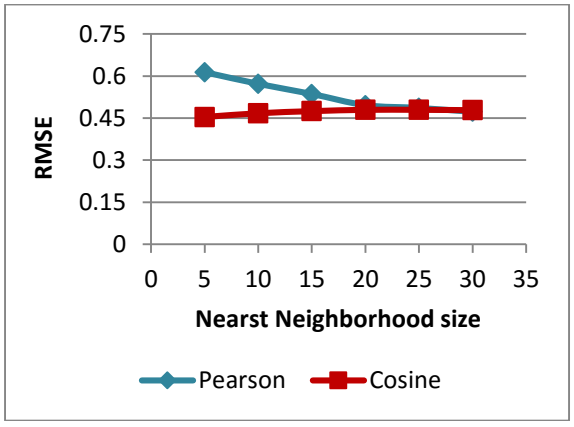


Fig. 4: Impact of neighborhood size on MAE and RMSE

As Fig. 4 shows, Pearson based and cosine based models show different types of sensitivity .In Pearson based model, as the neighborhood's size is increased the error decreases (prediction quality increases) and when this value reaches 30, the error will be minimized. Therefore, we can say that the optimal size of neighborhood is 30.

On the other hand, about cosine-based model, the prediction quality decreases by increasing the neighborhood size. Based on this observation, we select k=5 as optimal value for cosine based model.

Another important factor that affects the prediction quality is similarity measure. This study has used Pearson correlation and cosine similarity to find the similar users. For each similarity measure, we implemented the proposed algorithm to generate the prediction. Fig. 5 shows the results of the two different similarity measure on a given test set. Results shows that cosine similarity measure has better performance than Pearson similarity. Therefore, we chose cosine similarity for the rest of our experiments.

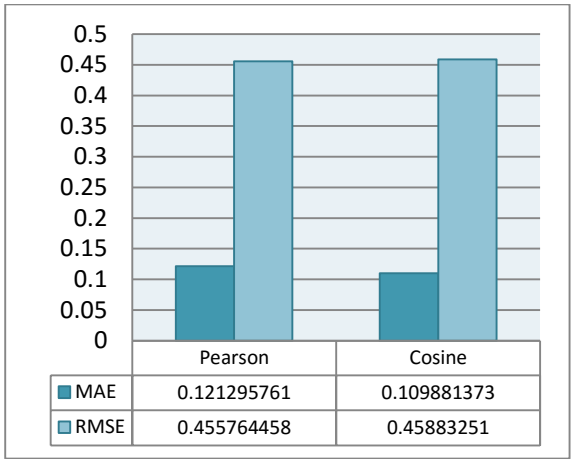


Fig. 5: Impact of the similarity computation measure on cluster based CF

We also surveyed the quality of the produced recommendations by using recall, precision and F-measure measures. Figure 6 shows the quality of recommendations with respect to different recommended locations. Figure 6 clearly specifies that precision has a reverse relationship with the number of recommended locations. So, precision decreases by increasing the number of recommended locations. On the other hand, recall has a direct relationship with the number of recommended locations. Thus, recall is increased by increasing the number of recommended locations.

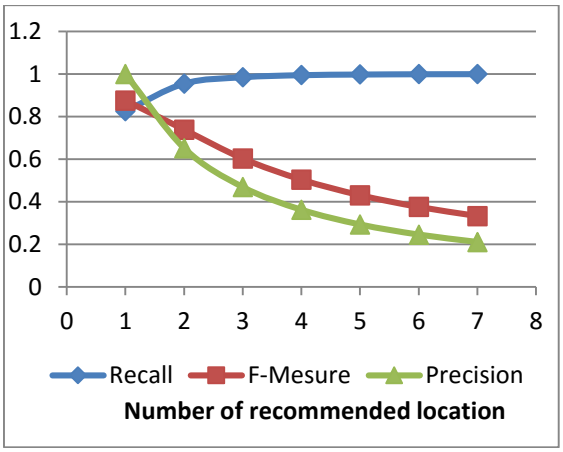


Fig. 6: Decision support measure comparison of mission location recommender system

Our purpose is to provide a ranked list of 3 recommendations. The results of observing recommendation evaluation metrics for 3 recommendations are shown in Table1. According to Table 1 we observed that considering the number of recommendations, precision metric has an acceptable rate. This means that most of the missioners can benefit from at least two of these three provided recommendations. Also high recall rate means that the

system is capable of recommending most of the locations that the missioner is interested in.

Table 1: Comparison of Recall, Precision, F-measure for 3 top recommendations.

Precision (%)	*Recall (%)*	*F-Measure (%)*
47.1	98.6	60.6

Finally, to compare the performance of the proposed cluster-based recommendation algorithm with the performance achieved by user-based algorithm (memory based), we performed an experiment which calculated user-based recommendation algorithms with optimal neighborhood size of 50 and used cosine measure to compute similarity between users.

We also compared the proposed method with the clustering method which is a model based method. In clustering method three locations which have the highest scores among that cluster's members are recommended to that cluster's users.

These results are shown in Table 2. We observe that cluster based CF outperforms user based CF by the tradeoff made between Recall, Precision and F-Measure.

Although recall in the model based method is better than cluster based CF, but as it is shown in Table 2, this method operates poorly according to precision and F-measure.

Altogether, as expected, we found that cluster based CF performs better than memory based and model based techniques.

Table 2: results of comparing the performance of different algorithms.

Evaluation metric / *Method*	*Recall (%)*	*Precision (%)*	*F-Measure (%)*
Memory Based (User Based CF)	97	45.4	59
Model Based (Clustering)	100	4.9	9.2
(hybrid) Cluster Based CF	98	47.1	60.6

6. CONCLUSION AND FUTURE WORK

This paper proposes an intelligent system that can help both the missioners in selection of the suited mission location, and the dispatching manager in allocating locations to missioner and improvement of dispatching programs. This study suggests recommender systems as the suitable solution for the mentioned goals.

As collaborative filtering is a common and successful method, this paper has also used this method to recommend location to missioners. By increase in the number of missioners traditional collaborative filtering will not be able to solve the scalability problem. Therefore, we have used cluster based CF. Our experimental results proved suitable performance of this approach. We also showed that cluster based CF is more accurate and scalable than user based CF and clustering technique.

In future, we plan to use user profile's data to overcome the new user problem. Additionally, we plan to engage content-based algorithm which takes into account the content of location, to improve the quality of further recommendations.

Acknowledgments

The authors would like to thank Mr. Saadatmand for providing the experimental data and helping with data realization.

Reference

[1] A. Kohrs, and B. Merialdo, "Clustering for Collaborative Filtering Applications". In Proceedings of CIMCA'99. IOS Press, (1999).

[2] A.M. Rashid, S.K. Lam, G. Karypis, and J. Riedl, "ClustKNN: A Highly Scalable Hybrid Model- & Memory-Based CF Algorithm". WEBKDD, (2006).

[3] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering", the Fifth International Conference on Computer and Information Technology (ICCIT), (2002).

[4] D. Bridge, and J. Kelleher, "Experiments in sparsity reduction: Using clustering in collaborative recommenders", in Procs. Of the Thirteenth Irish Conference on Artificial Intelligence and Cognitive Science, (2002), pp. 144–149.

[5] D. M. Pennock, E. Horvitz, S. Lawrence, C. L. Giles, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach", in Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI), (2000).

[6] E. Vozalis, and K. G. Margaritis, "Analysis of Recommender Systems' Algorithms", In The Sixth Hellenic European Conference on Computer Mathematics and its Applications (HERCMA 2003), (2003).

[7] G. Adomavicius and A Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions", IEEE Trans. Knowl. Data Eng, (2005), Vol.17, No. 6, pp. 734-749.

[8] G. Xue, C. Lin, and Q. Yang, "Scalable collaborative filtering using cluster-based smoothing". In Proceedings of the ACM SIGIR Conference, (2005), pp.114-121.

[9] H. Jon, A. K. Joseph, R. John, "An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms", (2002) Information Retrieval, Vol. 5, pp. 287-310.

[10] J. Herlocker, J. L. Konstan, G. Tervin and J. Riedl, "Evaluating collaborative filtering recommender systems". ACM Transactions on Information Systems, (2004), Vol. 22, No.1, pp. 5-53.

[11] J. Kelleher and D. Bridge, "Rectree centroid: An accurate, calable collaborative recommender". In Procs. Of the Fourteenth Irish Conference on Artificial Intelligence and Cognitive Science, (2003), pp. 89–94.

[12] M. C. Pham, Y. Cao, R. Klamma, M. Jarke, "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis", Journal of Universal Computer Science, (2011) , Vol. 17, No.4, pp. 583-604.

[13] M. Papagelis, D. Plexousakis, "Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents", Engineering Application of Artificial Intelligence, (2005), Vol.18, pp. 781-789.

[14] N. Mittal, R. Nayak, MC Govil and KC Jain, "Recommender System Framework using Clustering and Collaborative Filtering". 978-0-7695-4246-1/10- IEEE, Third International Conference on Emerging Trends in Engineering and Technology, (2010).

[15] Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen, "Scalable Collaborative Filtering Using Cluster-based Smoothing", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, (2005), Vol. 28,pp. 114-121.

[16] R. Jin and L. Si , "A Study of Methods for Normalizing User Ratings in Collaborative Filtering", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, (2004), pp.568-569.

[17] S. Gong, "A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering", Journal Of Software, (2010), VOL. 5, NO. 7.

[18] S.H.S. Chee, J. Han and K. Wang, "Rectree: An efficient collaborative filtering method", Lecture Notes in Computer Science 2114, (2001).

[19] SPSS INC. , Clementine ®12.0 Algorithms Guide, available from: http://www.spss.com

[20] Y. L. Chen, and L. C. Cheng, "A novel collaborative filtering approach for recommending ranked items". Expert Systems with Applications, (2008), Vol. 34, pp. 2396-2405.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A Context-based Prototype for decision making in database administration

Hassane TAHIR

Energy & Utilities Department, Sopra Steria Group
11 avenue du Maréchal Juin, Meudon-la-Forêt, 92366 Cedex, France
hassan.tahir@soprasteria.com, hassanetahir@hotmail.com

Abstract

Decision Support Systems (DSS) have a great role in assisting decision makers in many organizations to identify and solve problems in order to make decisions. In the area of database management, many approaches have been used to automate procedures set for complex activities such as performance and database recovery. However, procedures need to be contextualized in order to take into account the permanent changing of technical and social contextual elements added in DBA (Database Administrator) practices. This paper presents a context-based prototype for decision making to support experts in database management and administration. The prototype uses a software-modeling tool called Contextual Graphs (CxG).

Keywords: *Contextual Elements, Contextual Graphs, Database Administration, DBA, Decision Support System, Intelligent Assistant System, Practices, Procedures, Prototype for decision making.*

1. Introduction

Today with the fast evolution of new technologies as big data, social networks, cloud computing and mobile systems, the decision-making in organizations is becoming more and more complex. As a result, decision makers have been obliged to make the best decisions in the shortest possible time. In the area of database administration, support is needed for DBAs (Database Administrators) to make decisions about complex activities such as tuning problems and managing the continuous changes to databases. DBMS vendors continue to provide standard procedures for solving most of the incidents that have been well known for a long time (bad memory configuration, buffer caches, database crashes, and security bugs, among others). In addition, organizations have also established, from their perspectives, their own internal procedures for incident solving based on the basis of their experience. However, each DBA develops his own practice to solve an incident, and one observes almost as many practices as DBAs for a given procedure because each DBA tailors the procedure in order to take into account the current proceduralized context, which is particular and specific. In many working processes human beings can be observed to develop accurate procedures to reach the efficiency that decision makers intended when designing the task.

In parallel, DBAs prefer to plan again their action in real time rather than to rely on these procedures based on company's experience. This is due to two main reasons. Firstly, the selected procedure is not always perfectly adapted to the situation at hand and can lead to improper actions or sub-optimal incident resolution strategies. Secondly, if the DBA relies on a procedure, he can miss some important facts and notice them too late to adequately solve the incident. DBAs choose generally to plan again their action continuously according to the situation. Procedures are then used as frames to build and create a genuine strategy tailored to the specificity of a given situation. Such practices are based on operational knowledge and are shared by actors. In many different domain areas (i.e. medicine, technical process regulation, nuclear power, etc.…), the distinction between procedure and practice in the one hand, and the notion of context in the other hand is very important. Practices can appear as a contextualized expression of procedures.

The modeling of DBAs' reasoning is a difficult task because each DBA can use a number of contextual elements, and also because procedures for solving complex incidents do not always offer a great flexibility and degree of freedom. Their reasoning stems from some chunks of implicit knowledge, which are imposed on the DBA because they correspond to mandatory procedures. Procedure are established from DBA's experience during similar incidents and fixed by the company.

This paper presents a prototype for designing a context-based intelligent assistant system to support experts in database management and administration activities. The prototype uses a software-modeling tool called Contextual Graphs (CxG), which is well adapted to represent user procedures and practices. First, some of the related approaches will be discussed. Then we present a description of the context-based prototype for decision making including the Contextual Graphs formalism,

prototype architecture and a case study. Finally, we conclude our work.

2. Related work

Intelligent assistance is one of the important active research fields within Artificial Intelligence (AI). This section reviews some of the important approaches for intelligent assistance in database management over the years. Many expert systems have been introduced to help in simplifying the process of database design and development like the Generalized Expert System for Database Design (GESDD) developed by [6]. Such experts systems are mainly targeted for experts in database design, but not for novice designers with little experience because they do not provide adequate facilities for novice database designers to use the system. Another limit of expert systems for database management is that they didn't adopt a user-centered approach and they did not consider context explicitly. Decision support systems (DSS) have also been used in database management. Palvia [11] presented an interactive DSS tool, which supports the database designer in this task by providing several heuristic optimization procedures to enable the generation of many good designs. In addition, Spiegler and Widder [13] proposed a conceptual model based on a decision support system (DSS) that focuses mainly on the later stages of the system design, when the system is being mapped into the structure of the database management system (DBMS) selected for application. One of the main problems with such DSSs is that they didn't always consider, detect and process the users' context, preferences and new unexpected situations (i.e. DBA is new to the database administration tool). Other related approaches are Intelligent Tutoring Systems (ITS) for learning database management. These ITSs focus on teaching database domains such as Structured Query Language (SQL) and Database Design. One of the ITS is SQL LightweightTutoring Module (SQL-LTM) [7] which is a system that can provide semantic feedback on SQL statements, pointing out their logic flows, even if they are syntactically correct. Another work is that presented by Risco and Reye [12] about an evaluation of the Personal Access Tutor (PAT), an Intelligent Tutoring System (ITS) for Learning Rapid Application Development (RAD) in a database environment. Many other research efforts have already been made towards a best automatic database management and administration using multi-agent systems such as AutonomousDB tool proposed by Moraes et al. [9] to support the task of schema evolution in heterogeneous multi-database environments where there are replicated schemas. Other two famous multi-agent systems are Intelligent Agent Assistant (IAA) by Elfayoumy and Patel [8] and Grid Control Agent [10] to help DBAs in performance monitoring tasks and the automation of resolution actions.

In this section we have cited different approaches to intelligent assistance for supporting database management including database administration. Most of these approaches consider only technical sensors as contextual parameters and there is a little research dealing with context in the area of database and data administration from a user-centered perspective (i.e. DBA viewpoint). In information retrieval systems, Bouramoul et al. [1] proposed an approach based on the context to evaluate the performance of the search tool and the relevance of results compared to an executed query and the user's judgments. Another interesting approach is to explore system-level provenance to improve the mental models, and troubleshooting process for system administrators as in [5]. Nevertheless, context is rarely considered explicitly in procedures and policies used by many database administration tools. In addition, most policies rely on technical sensors, location and history. Contextual information does not include user's individual and interaction context (skills, experience, age, etc.…) with the system and other actors, and which could be of a great importance in efficiently performing complex DBA activities.

We can summarize the limitations of these approaches by saying that the developed tools are:

- Unable to automatically detect, diagnose and repair efficiently failures in unexpected new situations (context is evolved) ;
- Context-Aware Administration (i.e. only physical parameters and sensors are considered);
- Not Human-Centered Context (i.e. Social Context: DBA Profile, experience, Knowledge, Conflict with Developers, degree of collaboration between DBA and other shareholders…).
- Not suitable for the efficient sharing of context (i.e. Context is implicit).

3. Prototype for decision-making

This section presents a brief presentation of the Contextual Graph formalism followed by the architecture of the prototype for decision-making and a case study in the area of database administration.

3.1 Contextual Graphs formalism

The proposed context-based prototype for decision making uses a Contextual Graph formalism to represent the different ways to solve a problem. Each path corresponds to a practice, a way to fix the problem. It is a directed graph, acyclic with one input and one output and a general structure of spindle. Fig. 1 shows the elements in a Contextual Graph. This formalism and its implementation are well explained in [3]. Elements of a Contextual Graph are: actions, contextual elements, activities and temporal branching. Brézillon and Pomerol [4] consider that context is "what constrains something without intervening in it explicitly."

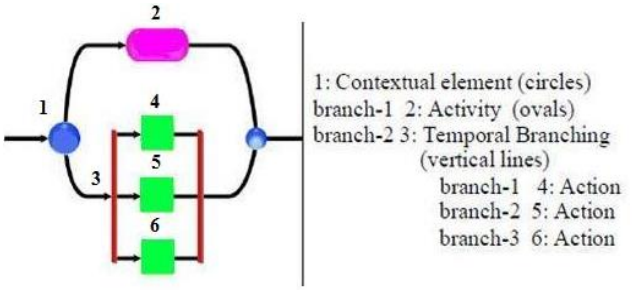


Fig. 1 Contextual Graph Elements

According to Brezillon [2], an Intelligent Assistant System (IAS) must present different properties like:

- Providing users with a first approximation of environmental trends and events;
- Pointing out useful information implicit in large volumes of data to alert users to sudden changes;
- Developing multiple scenarios and perspectives on a given line of action;
- Attracting user attention to existing and emerging strategic issues;
- Supporting users in sharing and communicating their views and perspectives;
- Guiding user attention to specific data and its interpretation in relation to particular issues;

The design of a Context-Based Intelligent Assistant System uses Contextual Graphs formalism to respond to the main requirements in the field of database administration such as:

- Assisting DBA in executing and managing efficiently their daily activities as well as resolving incidents;
- Improving collaboration between actors by sharing the context in which the DBA is confronted when dealing with complex database administration activities;
- Analyzing practices and the different strategies used by different actors when dealing with the same or similar situations and contexts.

Contextual Graphs represent the set of known practices (strategies) in order to solve a given problem. They also allow incremental acquisition of practices and provide an understandable way to model context-based reasoning. A practice is the path from the input to the output of a Contextual Graph. The problem solving process is guided throw a specific path by the evolution of context over time. Adopting a given practice or strategy among the others is dictated by the values of the different contextual elements forming the situation. However, it is not always obvious for a user to select one of these values.
User practices may differ from each other because of their contexts that are slightly different where users used different actions at a step of the problem solving. The process of practice acquisition by the CxG system concerns the new action to integrate and the contextual element that discriminates that action with the previous one. The integration of the new practice requires either adding a new branch on an existing contextual node, or introducing a new contextual node to distinguish the alternatives. The phase of incremental acquisition of practices relies on interaction between the CxG system and the users in order to acquire their expertise, which consists of a context-based strategy and its evolution along the process of the problem solving. We can distinguish two types of practices: (1) Practices created by experts (Design mode) (2) Practices executed by users (Running mode).

3.2 Prototype Architecture

As shown in Fig. 2, our proposed prototype for decision making integrates the following components:
CxG Editor: This component helps authorized users to manage their corresponding Contextual Graphs representing the main procedures and the significant changes added by them (i.e. practices).

CxG Reader: This component enables reading a desired Contextual Graph to execute one or more practices already created by different experts to perform a given activity.
CxG Analyzer: This component supports users in adopting the best strategies and practices when performing complex tasks to reach a desired goal.
CxG Manager: The CxG Manager controls and communicates with the different components of the Contextual Graphs Platform and with users.
Operational experience database: The CxG Manager uses this component to record and store users' practices.
Archive database: This component manages copies of executed Contextual Graphs.

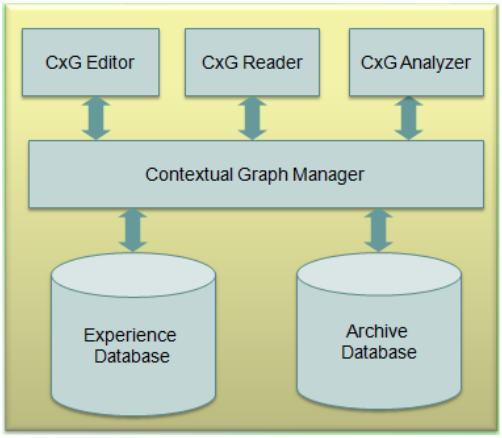


Fig. 2 Architecture of the proposed context-based prototype for decision making

3.3 A case study

In this case study, we have used the Contextual Graph representing a part of a DBA procedure for database performance troubleshooting as in [14] and [15]. To solve a serious performance problem within a given critical situation and context, a DBA may have different options when asking this question: what causes the slow response time of the system? Is it a network problem? Is it a bad database configuration? Is it a bad query in the application programs? Etc.
The Contextual Graph in Fig. 3 is composed of the following:

- Three contextual elements C1, C2 and C3 representing respectively nodes numbered 3, 7 and 11 with the set of values: Val (C1)= {C1.0, C1.1, C1.2}, Val (C2)= {C2.0, C2.1, C2.2}, Val (C3)= {C3.0, C3.1, C3.2, C3.3}.
- Set of Actions A= {A1, A2, A4, A5, A6, A7, A8, A9, A10, A12, A13, A14, A15}.

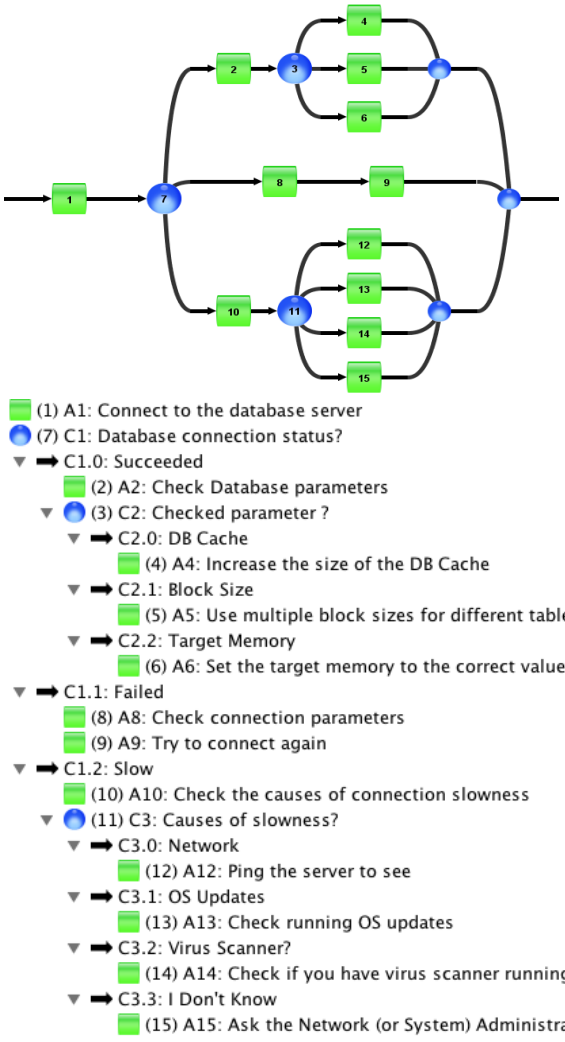


Fig. 3 Contextual Graph representing a part of a DBA performance procedure.

The DBA may be interested in the statistics about the path he selected, the number of errors generated when following that path but also the most used path for solving a critical problem within a context similar to that of his current situation. Many indicators can be used to assist the users of Contextual Graphs (i.e. the average evaluation of a selected contextual element branch, profile of user, number of executions, total time of task execution, etc.). Table 1 shows an example of the indicators about a selected branch of a contextual element.

Table 1: Contextual Elements Indicators

Contextual Element (CE)	*CE values*	*User Profile*	*Evaluation (%)*
Database connection status?	Succeeded	DBA	90
	Failed	DEVELOPER	60
	Slow	DBA	30
Checked parameter?	DB Cache	DBA	50
	Block Size	DBA	20
	Target Memory	SYSTEM ADMIN	10
		DBA	40
Causes of Slowness?	Network	DBA	10
		SYSTEM ADMIN	30
	OS Updates	SYSTEM ADMIN	40
	Virus Scanner	SYSTEM ADMIN	30
	I Don't Know	DBA	15

4. Conclusion

This paper has presented a context-based prototype for decision making to support database administrators (DBAs) in their complex activities such as resolving performance problems to ensure high availability of information systems. We have used the Contextual Graphs formalism to design the prototype. The objective is to build a context-based intelligent assistant system that can be used in different domains. Detailed architecture and evaluation of the described prototype will be explained in our future publications. We will also discuss how Data Marts can be used with Contextual Graphs to help in the process of analysis in decision making both in individual and collaborative activities.

References

[1] A. Bouramoul, M.-K. Kholladi, and B.-L Doan, Using Context to Improve the Evaluation of Information Retrieval Systems. International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.

[2] P. Brézillon, From expert systems to context-based intelligent assistant systems : a testimony. The Knowledge Engineering Review, 26(1) : 19-24, 2011.

[3] P. Brézillon, Task-realization models in Contextual Graphs. Modeling and Using Context (CONTEXT-05), A. Dey, B.Kokinov, D.Leake, R.Turner (Eds.), Springer Verlag, LNAI 3554, pp. 55-68, 2005

[4] P. Brézillon, and J.-C. Pomerol, Contextual knowledge and proceduralized context. In Proceedings of the AAAI-99 Workshop on Modeling Context in AI Applications, Orlando, Florida, USA, pages 16–20, 1999.

[5] M. Chiarini, Provenance for System Troubleshooting, http://static.usenix.org/event/tapp11/tech/final_files/Chiarini.pdf, Workshop on the Theory and Practice of Provenance (TaPP), Heraklion, Greece, June, 2011.

[6] A. Dogac, B. Yürüten, and S. Spaccapietra, “A Generalized Expert System for Database Design”, IEEE Transactions on Software Engineering, Volume 15, Issue 4, April, Page 479-491, 1989.

[7] R. Dollinger, Sql lightweight tutoring module - semantic analysis of sql queries based on xml representation and linq, in `Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2010', AACE, Toronto, Canada, pp. 3323-3328, 2010.

[8] S. Elfayoumy, and J. Patel, Database Performance Monitoring and Tuning Using Intelligent Agent Assistants. IKE 2012, in Hamid R. Arabnia, Leonidas Deligiannidis, Ray R. Hashemi Editors, WORLDCOMP’12, July 16-19, Las Vegas Nevada, USA, CSREA Press, 2012.

[9] A. C. Moraes, A. C. Salgado, and P. A. Tedesco, AutonomousDB: a Tool for Autonomic Propagation of Schema Updates in Heterogeneous Multi-Database Environments. IEEE, Fifth International Conference on Autonomic and Autonomous Systems, April, 20-25, pp. 251-256, 2009.

[10] Oracle, Grid Control Agent, 2013 Available at: http://www.oracle.com/technetwork/oem/grid-control/downloads/agentsoft-090381.html.

[11] P. Palvia, "An Interactive DSS Tool for Physical Database Design," Information Sciences, Vol. 54(3), April, pp. 239-262, 1991.

[12] S. Risco, and J. Reye, Evaluation of an Intelligent Tutoring System used for Teaching RAD in a Database Environment. Proceedings of the Fourteenth Australasian Computing Education Conference (ACE2012), Melbourne, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 123, Michael de Raadt and Angela Carbone, Ed. Australian Computer Society, Inc, 2012.

[13] I. Spiegler, and D. Widder, "Physical Database Design: A Decision Support Model", Data Base, Vol. 24,3 August, pp. 5-11, 1993.

[14] H. Tahir, and P. Brézillon, A Context-based approach for troubleshooting database problems. International Journal of Computer Science Issues (IJCSI), Volume 11, Issue 6, No 1, November 2014.

[15] H. Tahir, and P. Brézillon, Contextual graphs platform as a basis for designing a context-based intelligent assistant system. In: P. Brézillon, P. Blackburn, and R. Dapoigny (Eds.): CONTEXT 2013, LNAI 8175, pp. 259-273, 2013.

Hassane TAHIR has received a Higher Degree of Advanced Studies in Artificial Intelligence in 1998 and a PhD in Computer Science at UPMC in 2013 (University of Paris 6). Since 2008, he is working at Sopra Steria IT Company (France) as an Expert consultant in Oracle Database Management System and Business Intelligence. He is a Certified Professional from American Management Association (2007) and from Stanford University (2009). His research areas of interest are context modeling and management for decision-making, database management and administration, big data, computer security and business intelligence techniques particularly data warehousing.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Density Weighted Core Support Vector Machine

Shuxia Lu[1,*], Chenxu Zhu[2] and Caihong Jiao[1]

[1] Key Lab. of Machine Learning and Computational Intelligence,
College of Mathematics and Information Science, Hebei University
Baoding, Hebei 071002,China
cmclusx@126.com

[2] College of Science, Northwest Agriculture & Forestry University,
Yangling, Shanxi 712100, China
1441571065@qq.com

[1] Key Lab. of Machine Learning and Computational Intelligence,
College of Mathematics and Information Science, Hebei University
Baoding, Hebei 071002,China
1039877570@ qq.com

Abstract

Core Vector Machine (CVM) can be used to deal with large data sets classification problem, but CVM do not consider the density distribution of the data. In order to obtain the optimal description of the data, we propose a density weighted core support vector machine (DWCVM). In the proposed DWCVM, the relative density of each data point is based on the density distribution of the target data using the *k*-nearest neighbor (*k*-NN) approach. Experimental results on several benchmark data sets show that the performance of DWCVM is much better than CVM.

Keywords: *minimum enclosing ball, core set, support vector domain description, density, core vector machine.*

1. Introduction

Classification is a fundamental task in machine learning, data mining and pattern recognition. Prominent methods include support vector machine (SVM) [1], Kernel Density Estimator (KDE) [2], Support Vector Data Description (SVDD) [3], Small Sphere and Large Margin approach [4] for one-class classification and novelty detection, and so on. These methods involve solving the corresponding quadratic programming (QP) problems [5], which heavily limits the applicability of these methods for a large dataset.

In order to circumvent this drawback, many endeavors have been made to develop various techniques for scaling up these QP solvers. Typical techniques include chunking or some complicated decomposition methods such as the SMO algorithm [6]. Core Vector Machine (CVM) [7, 8] was proposed by Tsang et al. (2005), Tsang et al. proposed the core vector machine (CVM) by utilizing an approximation algorithm for the minimum enclosing ball (MEB) problem in computational geometry, the CVM algorithm achieves an asymptotic time complexity that is linear in N and a space complexity that is independent of N, where N is the size of the training patterns.

Inspired by [9] we propose a density weighted core support vector machine (DWCVM). In the proposed DWCVM, the relative density of each data point is based on the density distribution of the target data using the *k*-nearest neighbor (*k*-NN) approach. An optimal description of the data can be obtained by incorporating the weight into the search for using SVDD. Experimental results on several data sets demonstrate the effectiveness of DWCVM.

The rest of the paper is organized as follows. Section 2 reviews MEB, SVDD and GCVM. Section 3 describes the proposed DWCVM in detail. Experimental results are reported in Section 4. Concluding remarks are presented in Section 5.

2. Background

2.1 Standard MEB

The MEB problem aims to finding a smallest ball to enclose all training data defined by the sample set $S=\{x_i \mid x_i \in \mathbb{R}^n,\ i=1,\cdots,N\}$. The smallest ball denoted as $B(c,R)$ with center c and radius R . It is determined by solving

$$\min_{R,c} R^2 \quad s.t. \|\phi(x_i) - c\| \le R, \quad 1 \le i \le N. \tag{1}$$

which is similar to a one-class SVDD with hard margin and is called the standard MEB here. The corresponding dual of (1) is the following QP problem

$$\max_{\alpha} \boldsymbol{\alpha}^T diag(\mathbf{K}) - \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{K}\boldsymbol{\alpha} \quad s.t. \boldsymbol{\alpha}^T \mathbf{1} = 1, \quad 1 \le i \le N. \tag{2}$$

Where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2 \ldots, \alpha_N]^T \ge \mathbf{0}$ is the Lagrangian multipliers, $\mathbf{1} = [1,1,\ldots,1]^T$ is an *N*-dimensional vector, and $\mathbf{K} = [\phi(x_i)^T \phi(x_j)]_{N\times N} = [k(x_i, x_j)]_{N\times N}$ is the corresponding $N \times N$ kernel matrix with the term $k(x_i, x_j)$ denoting a kernel function.

2.2 SVDD

Tax and Duin (2004) presented the Support Vector Data Description (SVDD) which can obtain a spherically shaped boundary and the boundary that can be used to enclose normal data (similar to an enclosing sphere) and detect novel data or outliers (i.e. outside the enclosing sphere). The primal problem of SVDD is:

$$\min_{R,c,\xi_i} R^2 + C\sum_{i=1}^{N} \xi_i \quad s.t. \|\phi(x_i) - c\|^2 \le R^2 + \xi_i, \quad \forall i = 1,\ldots,N. \tag{3}$$

where *C* is regularized parameters which control the volume of boundary and the errors, and *c* and *R* respectively the center and radius of the sphere, denoted as $B(\mathrm{c}, \mathrm{R})$. The corresponding dual of (3) is

$$\max_{\alpha} \boldsymbol{\alpha}^T diag(\mathbf{K}) - \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{K}\boldsymbol{\alpha} \quad s.t. \boldsymbol{\alpha}^T \mathbf{1} = 1, \quad \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{C}. \tag{4}$$

where $\boldsymbol{\alpha} = \left[\alpha_1, \alpha_2, \ldots, \alpha_N\right]^T$ are the Lagrange multipliers.

2.3 The generalized core vector machine (GCVM）

The generalized core vector machine (The generalized CVM, GCVM) algorithm is proposed in [8]. The GCVM algorithm is much faster and can handle much larger datasets than existing SVM implementations. The generalized CVM algorithm can be used with any linear/nonlinear kernel and can also be applied to kernel methods such as SVR and the ranking SVM.

The GCVM utilizes an approximation algorithm for the center constrain minimum enclosing ball (CC-MEB) problem, which will be briefly introduced as follows:

The center and radius of a ball $B(\mathbf{c}, R)$ are denoted by $\mathbf{c}_B$ and r_B, respectively. Given an $\varepsilon > 0$, a ball $B(\mathbf{c}, (1+\varepsilon)R)$ is an $(1+\varepsilon)$ -approximation of $MEB(S)$ if $R \le r_{MEB(S)}$ and $S \subset B(\mathbf{c}, (1+\varepsilon)R)$. $\varphi : x_i \to \varphi(x_i)$ denotes the feature map associated with a given kernel *k*, and $B(\mathbf{c}, R)$ is the desired MEB in the kernel-induced feature space Γ.

The MEB problem finds the smallest ball containing all $\varphi(x_i) \in S$ in the feature space. In this section, we first augment an extra $\delta_i \in R$ to each $\varphi(x_i)$, forming $\begin{bmatrix} \varphi(x_i) \\ \delta_i \end{bmatrix}$. Then, we find the MEB for these augmented points, while at the same time constraining the last coordinate of the ball's center to be zero (i.e., of the form $\begin{bmatrix} \mathbf{c} \\ 0 \end{bmatrix}$). The primal form of the center constrain minimum enclosing ball (CC-MEB) problem can be formulated as

$$\min \quad R^2 \quad s.t. \quad \|\varphi(x_i) - \mathbf{c}\|^2 + \delta_i^2 \le R^2, \quad i = 1,\cdots,N. \tag{5}$$

The corresponding dual of (5) is the following QP problem

$$\max \quad \boldsymbol{\alpha}^T (diag(\mathbf{K}) + \boldsymbol{\Delta}) - \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha} \quad s.t. \quad \boldsymbol{\alpha}^T \mathbf{1} = 1, \quad \boldsymbol{\alpha} \ge \mathbf{0}. \tag{6}$$

where $K = [k(x_i, x_j)] = [\varphi(x_i)^T \varphi(x_j)]$ is the corresponding kernel matrix, and

$$\boldsymbol{\Delta} = [\delta_1^2, \cdots, \delta_N^2]^T \ge \mathbf{0}. \tag{7}$$

From the optimal $\boldsymbol{\alpha}$ solution of (6), we can recover *R* and $\mathbf{c}$ as

$$R = \sqrt{\boldsymbol{\alpha}^T (diag(\mathbf{K}) + \boldsymbol{\Delta}) - \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha}} \tag{8}$$

$$\mathbf{c} = \sum_{i=1}^{N} \alpha_i \varphi(x_i). \tag{9}$$

The squared distance between the center $\begin{bmatrix}\mathbf{c}\\0\end{bmatrix}$ and any point $\begin{bmatrix}\varphi(x_l)\\ \delta_l\end{bmatrix}$

$$\|\varphi(x_l)-\mathbf{c}\|^2+\delta_l^2=\|\mathbf{c}\|^2-2(\mathbf{K}\boldsymbol{\alpha})_l+k_{ll}+\delta_l^2. \quad (10)$$

which does not depend explicitly on the feature map φ.

Because of the constraint $\boldsymbol{\alpha}^T\mathbf{1}=1$ in (6), an arbitrary multiple of $\boldsymbol{\alpha}^T\mathbf{1}$ can be added to the objective without affecting its solution. In other words, for an arbitrary $\eta\in\square$, (6) yields the same optimal as

$$\begin{aligned}&\max \quad \boldsymbol{\alpha}^T(diag(\mathbf{K})+\boldsymbol{\Delta}-\eta\mathbf{1})-\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}\\ &s.t. \quad \boldsymbol{\alpha}^T\mathbf{1}=1, \quad \boldsymbol{\alpha}\geq\mathbf{0}.\end{aligned} \quad (11)$$

Hence, any QP problem of the form (11), with the condition (7), can also be regarded as a special MEB problem, called center constrained MEB, i.e. CC-MEB. As pointed out by Tsang et al., CC-MEB can be approximately solved with the asymptotic linear time complexity $O(N)$ and its space complexity independent of N for large datasets by using the generalized core vector machine.

The GCVM algorithm is introduced as follows:

The GCVM algorithm is shown in Algorithm 1. Here, the core set, the ball's center, and radius at the tth iteration are denoted by S_t, $\mathbf{c}_t$, and R_t respectively. The GCVM algorithm requires the input of a termination parameter ε.

Algorithm 1. GCVM

1) Initialize ε, $t=0, S_t, \mathbf{c}_t, R_t$.
2) Update the core set: if there is no training pattern that falls outside the ball $B(\mathbf{c}_t,(1+\varepsilon)R_t)$ in the corresponding feature space, $S=S_t$.
3) Find $\mathbf{z}$ such that it is the farthest away from $\mathbf{c}_t$ in the corresponding feature space and set $S_{t+1}=S_t\bigcup\{\mathbf{z}\}$.
4) Find the new MEB: $B(\mathbf{c}_{t+1},R_{t+1})$.
5) Set $t=t+1$, and go to step 2.

3. Density weighted core support vector machine

To accurately reflect the characteristics of the target data set, we propose a density weighted core support vector machine (DWCVM). In the proposed DWCVM, the relative density of each data point is based on the density distribution of the target data using the *k*-nearest neighbor (*k*-NN) approach. The distance between x_i and the *k*th nearest neighbor of x_i is denoted as $d(x_i,x_i^k)$; where x_i^k is the *k*th nearest neighbor of data point x_i. Using *k*-NN distance, the density weight of data point x_i is defined as:

$$w_i=1-\frac{d(x_i,x_i^k)}{\max_{j\in train\ set} d(x_j,x_j^k)} \quad (12)$$

Density weight measures the relative density based on the density distribution of the target data by comparing the *k*-NN distance of each data point with the maximum *k*-NN distance of the dataset. Density weight falls within the range $0\leq w_i\leq 1$.

To measure the density weight in feature space, can use the kernel function to map data into high dimensional space. The distribution of the data in feature space may be different from the original data distribution. In order to obtain a more appropriate description, we estimate the density weight in real space.

According to the density weight estimation method in (12), a data point located in a comparatively high-density area is close to its neighbors, so the distance between that data point and its *k*th nearest neighbor decreases, and eventually the density weight will become larger. In relatively low-density areas, data points are far from each other, so the density weight value will be low.

To apply the density weight, the objective function is defined as follows:

$$\begin{aligned}&\min_{R,c,\xi_i} R^2+C\sum_{i=1}^{N}w_i\xi_i\\ &s.t.\ \|\phi(x_i)-c\|^2\leq R^2+\xi_i, \quad \forall i=1,\ldots,N.\end{aligned} \quad (13)$$

We impose the weight w_i on each data point x_i. The data points in high-density regions receive a larger weight, so the effect of the slack variable is compounded. Therefore, to minimize the objective function, the spherical description will shift toward the high-density regions. On the other hand, with decreasing weight in relatively sparse areas, the influence of each data point will be reduced and

there is no pressure to keep data lying outside the spherical description.

By introducing the Lagrangian function for (13), and let partial differentiation of the Lagrangian function is equal to 0, we have the Wolf dual form

$$\begin{aligned} &\max_{\beta} \sum_{i=1}^{N} \beta_i \, \mathrm{k}(x_i, x_i) - \sum_{i,j=1}^{N} \beta_i \beta_j \, \mathrm{k}(x_i, x_j) \\ &s.t. \quad 0 \le \beta_i \le w_i C, \\ &\quad \sum_{i=1}^{N} \beta_i = 1, \quad \forall i = 1, 2, \ldots, N. \end{aligned} \tag{14}$$

Notice that, the upper bounds for Lagrange multipliers $\beta_i, i = 1, \ldots, N$ are no longer the same. Instead, each of them is respectively controlled by the corresponding weight. The primal variables can be recovered from the optimal β as

$$c = \sum_{i=1}^{N} \beta_i \phi(x_i), \quad R = \sqrt{\boldsymbol{\beta}^T diag(\mathbf{K}) - \boldsymbol{\beta}^{\mathbf{T}} \mathbf{K} \boldsymbol{\beta}}. \tag{15}$$

Therefore, by introducing density weight into the search for the optimal description of the dataset, can shift its description boundary to dense areas. In our proposed DWCVM, update core set by using density weight MEB method, and then train the core set by using SVM algorithm.

The DWCVM algorithm is introduced as follows:

Algorithm2. DWCVM

Step 1): Initialize $\varepsilon, t = 0, S_t, \mathbf{c}_t, R_t$.

Step 2): Update the core set: Terminate if there is no training point z such that $\phi(\mathrm{z})$ falls outside the $(1+\varepsilon)$ -ball $B(c_t, (1+\varepsilon)R_t)$ in the corresponding feature space, $S = S_t$.

Step 3): Find z such that $\phi(\mathrm{z})$ is furthest away from c_t in the corresponding feature space and set $S_{t+1} = S_t \bigcup \{z\}$.

Step 4): Find the new MEB: The c_{t+1} and R_{t+1} are computed by (15) .

Step 5): Set $t = t + 1$ and go back to step 2).

Step 6): Train the core set using SVM algorithm.

4. Experimental results

In this section, we compared the proposed algorithm DWCVM with CVM on several datasets for performance evaluation. In all experiments, the QP solver is adopted to solve the QP problem and the Gaussian function $k(x, y) = \exp(-\|x - y\|^2 / h)$ is taken as the kernel function, where h is the kernel parameter of the Gaussian kernel. In all experiments, the kernel parameter is $s^2/4$, s is the mean squared norm of the training data. All the experiments were carried out on a 3.1 GHz Pentium Core(TM) machine with 8GB RAM, running on the Matlab7.8 platform.

4.1 Data sets

The numbers of attributes, samples, positive samples and negative samples are shown in Table 1. The MiniBooNE dataset is used to distinguish electron neutrinos (signal) from muon neutrinos (background). The skin Segmentation dataset is constructed over B, G, R skin and Nonskin dataset is generated using skin textures from face images of diversity age, gender and race people.

We separately adopt the testing accuracy and geometric mean accuracy to evaluate the performance of algorithms. Considering the imbalanced nature of the training datasets, the geometric mean accuracy can be used. The geometric mean accuracy is defined as $g = \sqrt{a^+ \cdot a^-}$, where a^+ and a^- is computed by using Eq. (16). The measure takes into consideration the classification results on both positive and negative classes.

$$a^+ = \frac{\#\, positive\ sample\ correctly\ classified}{\#total\ \ positive\ sample\ classified} \times 100\%,$$

$$a^- = \frac{\#negative\ sample\ correctly\ classified}{\#total\ negative\ sample\ classified} \times 100\%. \tag{16}$$

Table 1: Summary of the data sets

Data sets	Attributes	Samples	Positive Samples	Negative Samples
MiniBooNE	51	130064	36499	93565
Spambase	58	4602	1813	2788
Skin	4	245057	50859	194198
Codrna	9	488565	162855	325710
Shuttle	10	58000	45586	12414
Sat	37	6435	3594	2841
Digit	65	5620	1697	3923

4.2 Performance evaluation

Experiment 1: In this experiment, we try to analyze the influence of the approximation parameter ε in the proposed DWCVM on the shuttle dataset. The percent 50

of samples are randomly selected as training data sets and the rest of the samples are used for testing data sets. The experimental results are listed in Table 2. From Table 2, we can see that as ε decreases, the geometric mean accuracy and the testing accuracy become higher, and the training time and the testing time become much more. Therefore, setting $\varepsilon = 1e-4$ is acceptable in the trade-off of the training speed and the classification accuracy for most cases.

Table 2: Influence of parameter ε on DWCVM

ε	g Accuracy	Testing Accuracy	Training Time (s)	Testing Time (s)
1e-2	82.54	85.21	0.11	1.24
1e-3	93.61	94.89	0.18	1.41
1e-4	94.25	97.12	0.25	1.74
1e-5	96.34	98.11	0.33	2.45
1e-6	97.21	98.46	0.71	3.21
1e-7	98.21	99.15	1.54	6.23

Table 3: Accuracy comparisons of DWCVM and CVM

Data sets	DWCVM		CVM	
	g Accuracy	Testing Accuracy	g Accuracy	Testing Accuracy
MiniBooNE	74.36	75.42	71.65	73.23
Spambase	75.64	76.24	75.25	74.10
Skin	98.57	98.87	95.45	92.62
Codrna	76.32	76.75	70.89	69.89
Shuttle	91.23	95.65	89.23	93.78
Sat	90.12	95.32	89.24	89.24
Digit	89.21	94.11	88.54	92.03

Table 4: Time comparisons of DWCVM and CVM

Data sets	DWCVM		CVM	
	Training Time (s)	Testing Time (s)	Training Time (s)	Testing Time (s)
MiniBooNE	28.03	20.11	13.10	16.56
Spambase	3.02	0.38	1.18	0.12
Skin	15.89	4.70	14.27	3.58
Codrna	27.32	4.65	37.84	9.25
Shuttle	1.56	0.20	1.71	0.21
Sat	8.33	0.80	4.31	1.23
Digit	31.10	2.01	42.54	2.22

Experiment 2: In this experiment, we compared the performance of DWCVM and CVM. For MiniBooNE and Skin Segmentation datasets, the percent 70 of samples are randomly selected as training data sets and the rest of samples are used for testing data sets. For the other data sets, the percent 50 of samples are randomly selected as training data sets and the rest of the samples are used for testing data sets. Table 3 and Table 4 illustrate the experimental results. The geometric accuracy and the testing accuracy comparisons of DWCVM and CVM are shown in Table 3. The training time and testing time comparisons of DWCVM and CVM are shown in Table 4. From Table 3, we can see that both the geometric accuracy and the testing accuracy of DWCVM are better than that of CVM. From Table 4, we can see that times of DWCVM and CVM are similar.

5. Conclusions

In order to consider the density distribution of the data, and deal with large data sets classification problem, we proposed the density weighted core support vector machine (DWCVM). In our proposed DWCVM, update core set by using density weight MEB method, and then train the core set by using SVM algorithm. The relative density of each data point is based on the density distribution of the target data using the *k*-nearest neighbor (*k*-NN) approach. Aims to accurately reflect the data density distribution of a target dataset with the weight of each data point based on relative density. This method prioritizes data points in high-density regions, and eventually the optimal description shifts to these regions. The application of a density measure for the dataset is beneficial for outlier detection, and generates a better performance. Experimental results on several data sets demonstrate the effectiveness of DWCVM.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (61170040), by the Natural Science Foundation of Hebei Province (F2015201185, F2013201220).

References

[1] C. C. Chang, and C. J. Lin, "Training v-support vector classifiers: theory and algorithms", Neural Computation, Vol.14, 2002, pp. 43-54.

[2] M. D. Marizio, and C. C. Taylor, "Kernel density classification and boosting: an L_2 analysis, Statistics and Computing", Vol. 15, No. 2, 2005, pp.13-123.

[3] D. M.J. Tax, and R. P. W. Duin, "Support vector data description", Machine Learning, Vol. 54, No. 1, 2004, pp: 45-66.

[4] M. R. Wu, and J. P. Ye, "A small sphere and large margin approach for novelty detection using training data with outlier", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 31, No. 11, 2009, pp. 2088-2092

[5] W. J. Hu, F. L. Chung, and S. T. Wang, "The maximum vector angular margin classifier and its fast training on large datasets using a core vector machine", Neural Networks, Vol. 27, 2012, pp. 60-73.

[6] N. Takahashi, and T. Nishi, "Rigorous proof of termination of SMO algorithm for support vector machines", IEEE Transaction on Neural Networks, Vol. 16, No. 3, 2005, pp. 774-776.
[7] I. W. Tsang, J. T. Kwok, and P. M. Cheung, "Core Vector Machine: Fast SVM training on very large data sets", Journal of Machine Learning Research, Vol. 6, 2005, pp. 363-392.
[8] I. W. Tsang, J. T. Kwok, and J. M. Zurada, "Generalized core vector machines", IEEE Transactions on Neural Networks, Vol. 17, No. 5, 2006, pp. 1126-1140.
[9] C. Myungraee, S. K. Jun, and B. Jun-Geol, "Density weighted support vector data description", Expert Systems with Applications, Vol. 41, 2014, pp. 3343–3350.

Shuxia Lu is a professor of the Faculty of Mathematics and Information Science, Hebei University. Received the B.Sc. and M.Sc. degrees in Mathematics from Hebei University, Baoding, China, in 1988 and 1991, respectively, and the Ph.D. degree from Hebei University, Baoding, China, in 2007. Her main research interests include machine learning and computational intelligence, SVMs.

Chenxu Zhu has been a B.Sc. degree candidate in College of Science from Northwest Agriculture & Forestry University, Yangling, China. Her research interests include computational intelligence.

Caihong Jiao received the M.Sc. degree in Applied Mathematics from Hebei University, Baoding, China, in 2015. Her research interests include Machine Learning.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Automatic gamma correction based on average of brightness

Pedram Babakhani[1], Parham Zarei[2]

[1]Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

Pe.Babakhani@sbu.ac.ir

[2]Department of Electrical Engineering, Bu-ali Sina University, Hamedan, Iran

Pzarei1993@gmail.com

Abstract

Preprocessing is essential stage in image processing because of limitation of imaging device or inappropriate environmental light. This paper presents a preprocessing technique for estimating the amount of gamma correction in the absence of any information or knowledge about environmental light and imaging device. The basic approach exploits the amount of gamma correction based on average brightness. The amount of gamma correction is then estimated by a power which transports average of brightness to center of histogram .

Keywords: *average of brightness, pre processing, histogram, gamma Correction, execution time*

1. Introduction

Luminance is an important factor in image processing that leads to perception of details. Technical limitation of imaging devices result non-linear effects on image. Gamma correction is non-linear operation which enhances brightness of image. Gamma correction is defined by the following power law expression:

$$S = T(R) = R^{\gamma}$$

S is the value of brightness in output image and R is value of brightness in original image that are mapped to [0 1].

If the value of gamma is known then inverting this process is obvious:

$$S = T(r) = r^{\frac{1}{\gamma}}$$

Gamma correction would be advantageous to remove non-linear effects in preprocessing stage For many applications in digital photography, image processing, and computer vision. In this paper a technique is presented for estimating the amount of gamma correction in the absence of any information or knowledge about environmental light or imaging device. The basic approach exploits the fact that amount of gamma correction is determined by transposing average of brightness to center of histogram.

2. Proposed method

Average of brightness is simple element that can be computed in the least amount of time. Basic approach in this article present a technique to estimate appropriate gamma based on average brightness. Although average of brightness doesn't present all information about image, it is the best choice to choose a sample between amounts of brightness in histogram. This method presents a technique which is different method and low order to estimate gamma. This paper proposes a method which estimates a power that transport average amount of brightness to center of histogram. This method extends the estimated power for gamma correction. This power can be chosen as global gamma for gamma correction. We suppose a gamma which changes average of brightness to $\frac{1}{2}$, then gamma is estimated based on following equations:

$$X^{\gamma} = \frac{1}{2}$$

$$\gamma = \log_X \frac{1}{2}$$

$$\gamma = \frac{\log_{10} \frac{1}{2}}{\log_{10} X}$$

$$\gamma = \frac{-0.3}{\log_{10} X}$$

X is average brightness and X ∈ [0 1]. In the equations, $\frac{1}{2}$ is center of histogram brighness which is global for any format and it is'nt limmited to Unint8 and int8 etc.

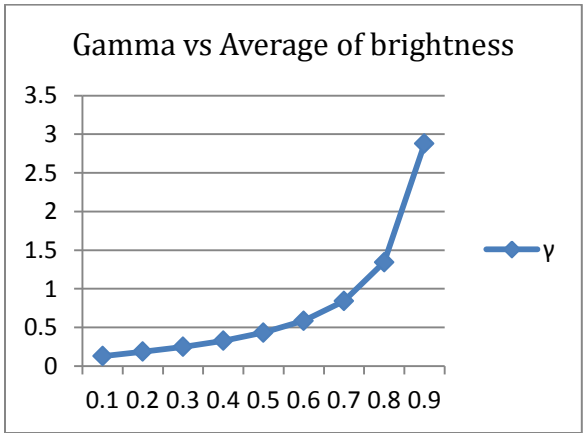


Fig. 1 Graph of Gamma Vs. Average of brightness

Finally, average of brightness in output image is not $\frac{1}{2}$ because this method just chooses the average of brightness in original image as a sample to estimate gamma. All pixels in output image will be enhanced with estimated gamma. After all input image will be enhanced with this method. Graph 1 demonstrates that proposed method estimates a logical and appropriate value for gamma correction.

3. Experimental results

In this paper we present a new preprocessing technique for estimating the gamma values without any information or knowledge of the imaging device or environmental luminance. We consider subjective and objective image quality assessment to demonstrate the performance of the proposed method. These figures are benchmark images with a high contrast and low contrast. The enhanced images bring out much more details of the original images. Quality of enhanced images indicates that the enhancement results using the proposed method have an appropriate performance compared to the other methods. In fact, our proposed method estimates gamma in least amount of time between existing approaches. Minimum execution time of proposed method is noticeable feature against other methods and algorithms.

Fig. 2 original image

Fig. 3 enhanced image

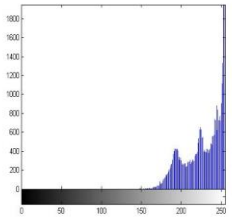


Fig. 4 original histogram

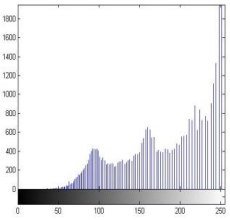


Fig. 5 enhanced histogram

Average of brightness in Figure 2 is 0.92 and estimated gamma is 3.5. more quality and more details (face and hat) in output image demonstrates that proposed method leads to enhancement.

Fig. 6 original image

Fig. 7 enhanced image

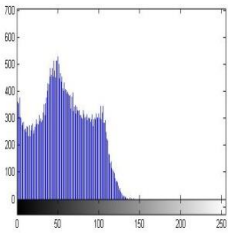


Fig. 8 original histogram

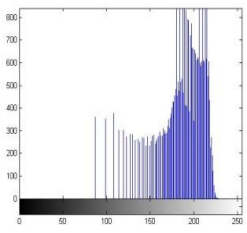


Fig. 9 enhanced histogram

Average of brightness in Figure 6 is 0.21 and estimated gamma is 0.2. more quality and more details (hair hand nose of baboon) proves that proposed method performs well.


Fig. 10 original image


Fig. 11 enhanced image

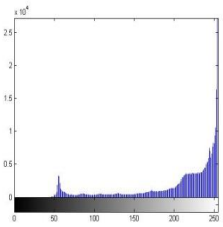


Fig. 12 original image

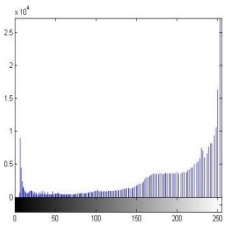


Fig. 13 enhanced image

Average of brightness in Figure 10 is 0.87 and estimated gamma is 2.33. more quality and more perceiving of details (reflection of object on mirror and details in shelf) shows that proposed method enhances original image.


Fig. 14 original image


Fig. 15 enhanced image

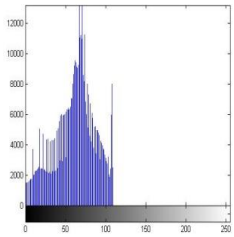


Fig. 16 original histogram

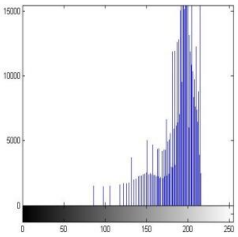


Fig. 17 enhanced histogram

Average of brightness in Figure 14 is 0.2.17 and estimated gamma is 0.19. more perceiving of details (houses and road) demonstrates that proposed method enhances image.
Histogram of all images became equalized and include more amount of brightness which leads to more contrast.

4. Comparison of results

Execution time is very important parameter in improvement of a method. Simulations demonstrate that Other methods such as local gamma correction and blind inverse gamma correction perform in more time than proposed method. All execution times are existed in Table 1.

Table 1 Execution times in different methods

Method	***Blind inverse gamma correction***	***Local gamma correction***	***Proposed method***
Execution time on 512X512	0.72 Sec	0.63 Sec	0.22 Sec
Execution time on 256X256	0.63 Sec	0.56 Sec	0.19 Sec
Execution time on 128X128	0.58 Sec	0.52 Sec	0.17 Sec
Execution time on 64X64	0.56 Sec	0.50 Sec	0.16 Sec
average	0.6225 Sec	0.5525 Sec	0.185 Sec

Execution time in proposed method is the least amount of time. Difference between proposed method and other methods is noticeable. Comparison between three mentioned methods are existed is Graph 1 and 2.

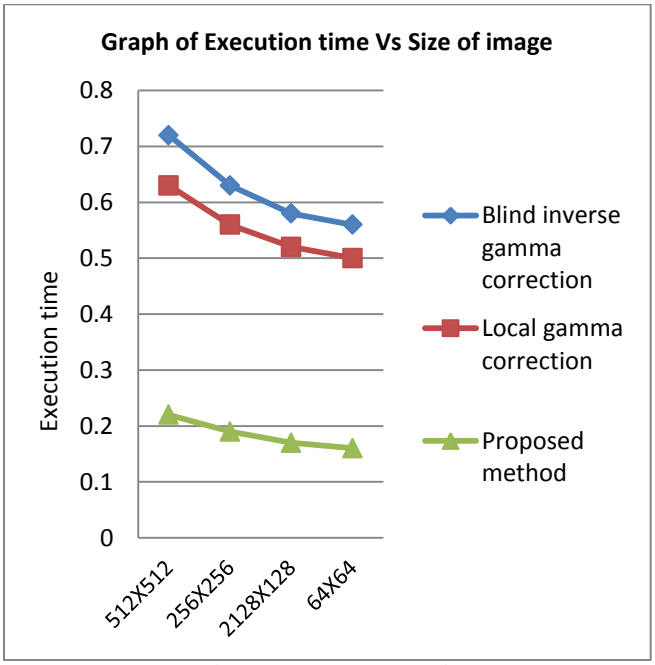


Fig. 18 Graph of Execution time Vs Size of image

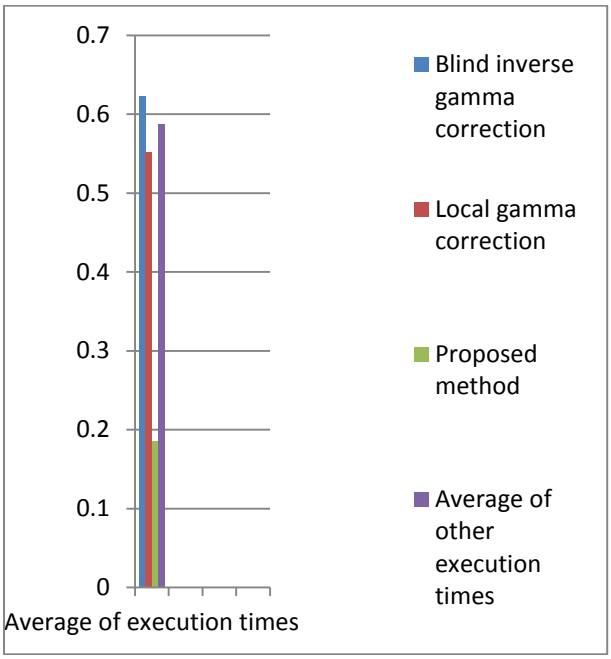


Fig. 19 Graph Average execution time Vs. Methods

Execution time is reduced from average 0.5875 second to 0.185 second (68.6 %) by proposed method. In sum up, proposed method performs at least time against other methods.

5. Conclusion

We have introduced a new image enhancement method based on gamma correction that estimates image gamma values without any calibration information or knowledge of the imaging device. The proposed method is a necessary preprocessing stage for most image analysis. Experimental results in this research indicate that the proposed method improves image quality, enhances the dynamic range and details of the image in least amount of time. On the other hand, proposed method performs in less time than other methods. This method can be implemented as a ASIC in the photography or printing devices.

References

[1] R.P. Kleihorst, R.L. Lagendiik, and J. Biemond. An adaptive order-statistic noise _lter for gammacorrected image sequences. *IEEE Transactions on Image Processing*, 6(10):1442.1446, 1997.

[2] Pizurica, A. and Philips, W. “Estimating the probability of the presence of a signal of interest in multiresolution single and multiband image denoising”, ***IEEE Trans. Image Process***, 654–665, 2006.

[3] Shi, Y., Yang, J. and Wu, R. "Reducing Illumination Based on Nonlinear Gamma Correction," ***In Proc.* 310** - Vol. 24, No. 4, 2011.

[4] Gonzalez, R. C. and Woods, R. E. Digital Image Processing. Prentice Hall, Upper Saddle River, NJ 07458, 2002.

[5] Farid, H. "Blind inverse gamma correction", ***IEEE Transactions on Image Processing***, Vol. 10, pp. 1428-1433, 2001.

[6] Pizer, S. M., E.P. Ambum and J.D. Austin, “Adaptive histogram equalization and its variation”, ***Computer Vision, Graphics, and Image Processing*** Vol. 39, No. 3, 355–368, 1987.

[7] Chen, S. D. and Ramli, A. R. “Minimum mean brightness error bihistogram equalization in contrast enhancement”, ***IEEE Transactions on Consumer Electronics,*** Vol. 49, No. 4, 1310–1319, 2003.

[8] FarshbafDoustar, M. and Hassanpour, H. "A Locally- Adaptive Approach For Image Gamma Correction," 10th International Conference on Information Sciences, ***Signal Processing and their Applications (ISSPA2010)*** 73-76. (2010)

[9] Asadi Amiri, S., Hassanpour ,H. and Pouyan, A.“Texture Based Image Enhancement Using Gamma Correction”, ***Middle-East Journal of Scientific Research*** , Vol. 6, 569-574. (2010)

[10] Guillon, S., Baylou, P., Najim, M. and Keskes, N. “Adaptive nonlinear filters for 2D a nd 3D image enhancement”, ***Signal Processing***, Vol. 67, No. 3 ,237–254, 1998.

[11] Babakhani, P. “automatic gamma correction based on average and deviation from center of histogram”. 2th International Conference on Advances in Engineering and Basic Sciences,76 – 82, 2015.

[12] Hasanpour, H. Asadi Amiri, S. Pouyan, A, A. “Automatic brighness enhancement based on local gamma correction”. National conference of computer engineering and information technology. Hamedan, 2011.

Pedram Babakhani is Bachelor Student in Department of Computer Science and Engineering in Shahid Beheshti University, Iran. His research interests include image and sound processing, hardware design and FPGA.

Parham Zarei is Bachelor Student in Department of Electrical Engineering in Bu-Ali University. Iran. His research inerests include signal and image processing, microwave and nonlinear RF.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Developing an Allocation Framework for Information Security Systems

Abdel Nasser H. Zaied[1], Walid I. Khedr[2] and Shimaa S. Mohamed[3]

[1] Dean, Faculty of Computer and Informatics, Zagazig University, Zagazig, Egypt
nasserhr@gmail.com

[2] Information Technology department, Faculty of Computer and Informatics, Zagazig University, Zagazig, Egypt
wkhedr@zu.edu.eg

[3] Decision Support department, Faculty of Computer and Informatics, Zagazig University, Zagazig, Egypt
Shimaa_said1100@yahoo.com

Abstract

Databases hold a critical concentration of sensitive information and become available on the internet to facilitate access, and as a result, databases are vulnerable and become the target of hackers. Today the security of database system become one of the most urgent tasks in database research, so to protect database system from attacking and compromised through authorized users who abuse or misuse data and unauthorized users who made unprivileged access. In this paper most of database vulnerabilities and threats which may face database system are reviewed and allocated proposed security techniques to protect database system from these threats to reduce risk of attacking database system.

Keywords: *Database Threats, Database Threats components, Security Techniques, Allocation Techniques, SQL Injection.*

1. Introduction

Data is most important and valuable asset in today's world as it helps organizations as well as individuals to extract information and use it to make various decisions and it is used in day-to-day life. Data are generally stored in database so that retrieving and maintaining it becomes easy, efficient and manageable. At a very general level, a database can be defined as a persistent collection of related data, where data are facts that have an implicit meaning. Typically, a database is built to store logically interrelated data representing some aspects of the real world, which must be collected, processed, and made accessible to a given user population. The database is constructed according to a data model which defines the way in which data and interrelationships between them can be represented [1]. As database have huge amount of sensitive information it is important to know challenges which face database system to protect it from attacking.

2. Database Threats and Vulnerabilities

With the increase in access to data and information stored in databases, the frequency of attacks against those databases has also increased. A database threat refers to an object, person or other entity that represents a risk of loss or corruption of sensitive data and sensitive information to an asset also database threats may be caused because of vulnerabilities in database system [2]. Attacks on database can also be classified into passive and active attacks [3]:

- **Passive Attack**: attacker only observes data present in the database. Passive attack can be done in following three ways (Static leakage, Linkage leakage, and Dynamic leakage).
- **Active Attacks**: actual database values are modified. These are more problematic than passive attacks because they can mislead a user, for example a user getting wrong information in result of a query. There are some ways of performing such kind of attack Spoofing, Splicing, and Replay.

Attacks on database can be used to disclose information, to sidestep authentication mechanisms, to alter the database, and to execute arbitrary code, in certain instances, on the database server itself. Attackers can be categorized into internal and external. External person is an intruder who gains access to a computer system and tries to infiltrates a database server to steal or tamper with data information. Internal, insider is an authorized user in database system that belongs to the group of trusted users and tries to get information that he is unauthorized to access, or administrator is a person who has privileges in administering a computer system, but abuses his rights and his power in order to extract valuable information; no database security can be guaranteed [4]. According to Ponemon Institute [5] in his study, the average total cost per data breach increased 15 percent to $3.5 million, and average cost per lost or stolen record increased more than 9 percent from $136 in 2013 to $145 in this year's study.

2.1 Database Threats and Vulnerabilities Components

In the period from 1999 to 2015, many researchers studied types of database threats and vulnerabilities; they classified them into 22 threats as follows:

2.1.1 SQL Injection Attack (SQLIA)

Imperva defined SQL injection attack as insertion (or "injection") of unauthorized SQL database statements into a vulnerable SQL data channel. SQLIA is considered anyone who can send untrusted data to the system. Typically, targeted data channels included stored procedures and Web application input parameters [6 & 7]. OWASP concluded that using SQL injection, attackers may gain unlimited access to a whole database and to the potentially sensitive information these databases contain [8].

The sources of SQL Injection can be one of the following [2]**:**

a. Injection through user input; malicious strings in web forms in web application.
b. Injection through cookies; modified cookie fields contain attack strings.
c. Injection through server variables; headers are be manipulated to contain attack strings.
d. Second-order injection; Trojan horse input seems fine until is used in a certain situation. Attacks don't occur when it first reaches the database, but when is used later on.

2.1.2 Weak Authentication

Amichai studied weak authentication as allowing the attackers to steal the identity of authorized database. An attacker may define any number of strategies to obtain credential. Weak authentication schemes allow attackers to assume the identity of legitimate database users by stealing or otherwise obtaining login credentials [9].
Approaches that an attacker may employ to obtain credentials as [9 & 10]:

- **Brute Force** – This approach attacker repeatedly enters all possible username/password combinations until he finds one that works.
- **Social Engineering** – This approach the attacker takes advantage the natural human tendency to trust in order to convince others to provide their login credentials.
- **Direct Credential Theft** – Here attacker may steal login credentials by copying post-it notes, password files, etc., Center's (ADC) ongoing research into proprietary database communication protocols and vulnerabilities.

2.1.3 Unmanaged Sensitive Data (Unauthorized Copies of Sensitive Data)

Many companies struggle to maintain an accurate inventory of their databases and the critical data objects contained within them. Forgotten databases may contain sensitive information, and new databases can emerge – e.g., in application testing environments – without visibility to the security team. Sensitive data in these databases will be exposed to threats if the required controls and permissions are not implemented [6 & 7]. Also, many web applications do not properly protect sensitive data, such as credit cards, tax IDs, and authentication credentials. Attackers may steal or modify such weakly protected data to conduct credit card fraud, identity theft, or other crimes [8].

2.1.4 Storage Media Exposure (Backup Data Exposure)

Backup storage media is often completely unprotected from attack (Unencrypted data on backup tapes and disk). As a result, numerous security breaches have involved the theft of database backup disks and tapes. Furthermore, failure to audit and monitor the activities of administrators who have low-level access to sensitive information can put your data at risk [9].

2.1.5 Web application attacks

Web application attacks through poorly configured websites, applications and databases. Today, the focus of exploitation has shifted from the operating system to the Web browser and multimedia applications. Web applications being used as the major platform for the flow of sensitive information there is increasing security concerns for the organizations as well as for the individuals. Due to transaction of high sensitive corporate information through the web and increase in online traffic multifold the security issue [11]. Web applications are vulnerable to a variety of well publicized attacks, such as cross-site scripting (XSS) and Cross-Site Request Forgery (CSRF) [12].

2.1.6 Buffer Overflow Attacks

A buffer overflow condition exists when a program attempts to put more data in a buffer than it can hold or when a program attempts to put data in a memory area past a buffer or unauthorized user causing the application to perform an action the application was not intended to perform. The overall goal of a buffer overflow attack is to subvert the function of a privileged program so that the attacker can take control of that program [13 & 14].

2.1.7 Advanced Persistent Threat (APT)

An advanced persistent threat (APT) is a kind of network attack in which an unauthorized person gains access to a network and stays there hidden for a long period of time. APT usually targets organizations and or nations for business or political motives. APT processes require high degree of covertness over a long period of time. As the name implies, APT consists of three major components/processes: advanced, persistent, and threat. The advanced process signifies sophisticated techniques using malware to exploit vulnerabilities in systems. The persistent process suggests that an external command and control is continuously monitoring and extracting data off a specific target. The threat process indicates human involvement in orchestrating the attack [15 & 16].

2.1.8 Covert Channel

Steven defined covert channel as means of communicating on a computer system, where both the sender and receiver collude to leak information, over a channel not intended for the communication taking place, in violation of a mandatory access control security policy and consider it as a computer security attack which can be used to weaken the system's security policy [17].

2.1.9 Unpatched DBMS

In database vulnerabilities are remain changing that can be exploited by unauthorized user, database suppliers release patches to ensure sensitive information in databases is protected from attackers. Once these patches are released they should be patched immediately. If left unpatched, hackers can reverse engineer the patch, or can often find information online on how to exploit the unpatched vulnerabilities, leaving a DBMS even more vulnerable that before the patch was released [18]. Attackers release unpatched vulnerabilities which can occur at any layer of a system which have sensitive information.

2.1.10 Redundant DBMS Features Enabled

There are many unnecessary features which are enabled by default in DBMS. And these unnecessary features should be turned off. If these unnecessary features are not change off so by this it can be dangerous attack on database. Attackers will only have more to use against you [18].

2.1.11 Broken Configuration Management (Misconfiguration)

Unwanted features are enabled in DBMS due wrong configuration. Incorrect or Unnecessary Implementation of Security at any Layer of a System Security misconfiguration can occur at any layer of a system. The user will provide unauthorized access or knowledge of a system for attackers [8].

2.1.12 Inference (Statistical Inference)

This is a database system technique which used to attack databases where malicious users gather sensitive information from complex databases at a high level. It is performed by analyzing number of different data sources in order to illegally get knowledge about a database. In basic terms, inference is a data mining technique used to predict and find information hidden from normal users. An inference presents a security breach if more highly classified information can be inferred from less classified information [19 & 20].
There are two inference vulnerabilities in database [3 & 20]:

Data Association: It occurs when two values have been taken together. And those are classified at a higher level than the classification of either value individually.

Data Aggregation: it occurs when a set of information is classified at a higher level than the individual level of data.

2.1.13 Social Engineering

Social engineering (known as non-technical or human-based attack) describes a method of launching attacks against information and information systems and targeting the existing vulnerabilities of both people and technology; as a result it is considered as the biggest security threat faced by both organization and individuals today. The types of information these attackers are seeking can vary, trying to ploy you into giving them your passwords or bank information, or access your computer to secretly install malicious software–that will give them control over your system [21].

2.1.14 Malware

Malware defined as software designed to attack and damage, disable, or disrupt computers, computer systems, or networks this mean that website malware can imagine, this makes website malware particularly insidious and dangerous. Malware includes Viruses, Worms, Spyware, Trojans, Bots, and other malicious programs. The reasons that make website vulnerabilities to malware, website owners continue to increase their website's popularity. Also increased interactivity on websites can introduce exploits that open the door to malware [22].

2.1.15 Database Rootkits

Rootkit is code or program or procedure run on a system by an intruder, or changes made to a system's internal state in order to retain control of key system resources without detection by user or administrator. Rootkits are often categorized as a variant of malware but differ in several important respects. The fundamental difference is scale of target – they're narrowly targeted, with a specific mission and capture and optionally modify specific data over a period of time without detection "Rootkits are used as a means of carrying out espionage". In order to install a rootkit, an attacker will require the ability to execute code on the target system. Furthermore, the attacker will need to run this code with administrative privilege, or exploit vulnerability in the operating system [23].

2.1.16 Excessive Privilege Abuse

When database users are provided with the access rights that allow them to perform other tasks not included in their job (users have privileges exceed their job requirement), these privileges may be abused purposely or accidentally, harmful intent can be discovered through such tasks thus leading to misuse of such privileges. For example, in a university administrator whose job requires only the ability to change student contact information may take advantage of excessive database update privileges to change marks [10 & 24].

2.1.17 Legitimate Privilege Abuse

Users will abuse legitimate database privileges for unauthorized purposes. When the authorized user misuses the authorized privilege for illegitimate purpose, this is the mean legitimate privilege abuse. For example a hypothetical rogue healthcare worker has privileges to view individual patient records via a custom Web application. The structure of the Web application normally limits users to view an individual patient's healthcare history – multiple records can't be viewed simultaneously and electronic copies are not allowed. However, the rogue worker can circumvent these limitations by connecting to the database using an alternative client such as MS-Excel and MS-SQL. Using MS-Excel, MS-SQL Server or Oracle and his legitimate login credentials, the worker may retrieve and save all patient records [6].

2.1.18 Privilege Elevation

Privilege Elevation Attackers may take advantage of database software vulnerabilities to discover flow of flaws which is taken advantage of by attackers and may result in the change of privileges such as converting access privileges from those of an ordinary user to those of an administrator. Vulnerabilities may be found in, built-in functions, stored procedures, protocols implementations, and even SQL statements. For example, a software developer at a financial institution might take advantage of a vulnerable function to gain the database administrative privilege. With administrative privilege, the rogue developer may turn off audit mechanisms, transfer funds, create bogus accounts, misinterpretation of certain sensitive analytical information, etc. [9 & 24].

2.1.19 Database Platform Vulnerabilities

Vulnerabilities in operating systems vulnerabilities and additional services installed on a database server could lead to leakage easily. Vulnerabilities in the previous operating systems such as Widows 98, Windows 2000, UNIX, etc. may lead to unauthorized access, data loss from a database, data corruption or service denial conditions. For example, the blaster worm created denial of service conditions from vulnerabilities which found in Windows 2000 [2].

2.1.20 Database Communication Protocol Vulnerabilities

Maximum numbers of security vulnerabilities are being identified in the database communication protocols of all database vendors. Four out of seven security fixes in the two most recent IBM DB2 FixPacks address protocol vulnerabilities1. Similarly, 11 out of 23 database vulnerabilities fixed in the most recent Oracle quarterly patch relate to protocols. Fake of activity targeting these vulnerabilities can range from unauthorized data access, to data corruption, to denial of service. For example, the SQL Slammer2 worm, took advantage of a flaw in the Microsoft SQL Server protocol to force denial of service and to carry out code on targeted database server [9 & 10].

2.1.21 Weak Database Audit Trail

Weak Database Audit Trial defined as automated recording of database transactions involving all sensitive data should be part of any database deployment and the database security considerations. Failure and absence (weak or non-existent) to collect detailed audit records of database activity may cause instability in operations and represents a serious organizational risk on many levels; such as regulatory risk, prevention, detection and recovery, this mean that audit policies that rely on built-in database mechanisms suffer a number of weaknesses that limit or preclude deployment [10 & 24].

2.1.22 Denial of Service (DOS) Attack

Denial of service (DOS) conditions could be created by many techniques which are related to the other mentioned vulnerabilities in database such as database platform vulnerabilities to crash database server. For example, attempts to "flood" a network, thereby preventing legitimate network traffic, attempts to disrupt connections between two machines, attempts to prevent particular individuals from accessing a service and attempts to disrupt service to a specific system. This attack is very serious attack [6 & 7].

3. Experimental Analysis

In this paper are made experimental analysis on previous 22 threats and vulnerabilities, these analysis was based on some features of database threats such as Attack type, Source of threats, Exploitability, Impact, and Users as shown in table (1). Attack type feature refer to Active or Passive attack all threats are active attack but inference threats, and when this attack happen which attack on database system (Database Server, Web Server, Browser, Network Infrastructure, and Operating System Attack). Source of threats feature refer to where this attack may happen such as Internal mean that this attack happened inside the system (Intra-organization-LAN Network), External mean that this attack happened outside the system (Extra-organization-WAN Network), and Internal - External mean that this threat may happen inside and outside system. Exploitability feature refers to ability users to hack system such as Easy, Average, Difficult, and Very Difficult. Easy means that freely available exploit code, exploit SQL Injection, Platform vulnerabilities, database vulnerabilities and there are easy to install malicious programs which download itself to users' computers without their knowledge such as malware. Average mean that attackers need administrative privilege, analyze number of different data sources in order to illegally get knowledge about a database, or need to know identity of authorized users before attack. Difficult mean privileged users who can access sensitive data. Very Difficult mean that attackers need to access network as privileged user and stay unknown for long time and target privileged users as in Advanced Persistent Threat (APT) threat. Impact feature refer to degree effect threat on system such as Severe and Moderate. Severe mean that the attackers can do anything the victim to obtain privileges of authorized users and reputation of organization could be harmed, Lack of accountability, Denial of service, Lead to complete host takeover. In this feature all data could be stolen, modified, or deleted and reputation of organization could be harmed also business impact of public exposure of the vulnerabilities, all accounts or some of them can be attacked as in SQL Injection, Weak Authentication, Unmanaged Sensitive Data, Backup Data Exposure, Legitimate Privilege Abuse, Database Platform Vulnerabilities, and Database Communications Protocol Vulnerabilities. Also may subvert the function of a privileged program, corrupt data, crash the program, and execute malicious code as in Buffer Overflow Attack. Attacker's goal in this feature may steal data rather than cause damage to the network or organization. APT attacks target organizations in sectors with high-sensitive information, for instance national defense, manufacturing and the financial industry. Malware attack cut corners with insufficient input validation on user input, inadequate logging mechanisms, and using fail-open error handling or failing to close a database connection. Penetrate organizations and steal sensitive data and including identity theft and financial ruin. Damage, disable, or disrupt computers, computer systems, or networks and loss your reputation, loss of customer trust and goodwill, downtime due to blacklisting and non-compliance issues violations. Rootkits may also modify the database object itself and change the execution path and switch off alerts triggered by Intrusion Prevention Systems (IPS) and modify a running operating system kernel in order to hide an attacker's presence. Not discovered after compromising a system. Excessive Privilege Abuse and Privilege Elevation in these attacks any "minor" breach becomes a major incident, gain DBA access (full control of the database), complete operating system control, and turn off audit mechanisms. DOS Attack cause data corruption, network flooding, resource consumption and resource server overload (memory, CPU, etc.), disrupt connections between two machines, prevent particular individuals from accessing a service and disrupt service to a specific system, crash database server (database is unavailable), paralyzing the entire operations of an organization or part of it. Moderate Features means that attackers can execute scripts in a victim's browser to hijack user sessions, deface web sites, insert hostile content, and redirect user's browser and impact to your reputation as in Web Application Attack. Also because of weak the system's security policies lead to leak sensitive information and financial losses and damage also reputation of organization is affected as in Cover Channel and Social Engineering Attack. Also all of data could be stolen or modified slowly over time, the system could be completely compromised without you knowing it, traffic, or full database take over, and recovery costs could be expensive as in Unpatched DBMS, Redundant DBMS Features Enabled, and Broken Configuration Management (Misconfiguration). All high classified data could be stolen, modified, or deleted, Could your reputation be harmed as in Inference Attack. Also Weak Database Audit Trial limits or precludes deployment such as Lack of User

Accountability, Performance Degradation, Separation of Duties, Limited Granularity and Proprietary.

Table 1: Database threats Features

Threats	Attack Type	Threats Source	Exploit	Impact
SQL Injection	Database Server Attack	Internal - External	Easy	Severe
Weak Authentication	Database Server Attack	Internal - External	Average	Severe
Unmanaged Sensitive Data	Database Server Attack	Internal	Difficult	Severe
Back up Data Exposure	Database Server Attack	Internal	Difficult	Severe
Web Application Attacks	Web Server Attack	External	Average	Moderate
Buffer Overflow Attack	Web Server Attack	External	Easy	Severe
Advanced Persistent Threat	Network Infrastructure Attack	Internal - External	Very Difficult	Severe
Covert Channel	Computer Security Attack	External	Average	Moderate
Unpatched DBMS	Database Server Attack	Internal	Easy	Moderate
Redundant DBMS Features Enabled	Database Server Attack	Internal	Easy	Moderate
Broken Configuration Management	Database Server Attack	Internal	Easy	Moderate
Inference	Database Server Attack	Internal - External	Average	Moderate
Social Engineering	Web Browser Attack	External	Difficult	Moderate
Malware	Web Server and Browser Attack	Internal - External	Easy	Severe
Database Rootkits	Web Server and Database Server Attack	Internal - External	Average	Severe
Excessive Privilege Abuse	Database Server Attack	Internal	Easy	Severe
Legitimate Privilege Abuse	Database Server Attack	Internal	Easy	Severe
Privilege Elevation	Database Server Attack	Internal	Difficult	Severe
Database Platform Vulnerabilities	Operating System Attack	Internal	Easy	Severe
Database Communication Protocol Vulnerabilities	Database Server Attack	Internal - External	Difficult	Severe
Weak Database Audit Trial	Database Server Attack	Internal	Difficult	Moderate
Denial of Service Attack	Database Server, Web Server and Network Infrastructure Attack	Internal - External	Easy	Severe

4. Proposed Security Techniques

The database attackers will gain money by selling sensitive information, which includes credit card numbers, Social Security Numbers, criminal records and important organization information etc. So, the need to insure the integrity of the data and secure the data from unintended access is emerged. To secure a database environment, many database security techniques are developed [2]. Database security depends on a set of systems, processes, roles, and procedures that can protect the database from unintended activities. Unintended activities can be categorized as authenticated misuse, malicious attacks or inadvertent mistakes made by authorized individuals or processes. The importance of database security will continue to grow as more data is shared, retained, transmitted and archived electronically [25]. After previous discussion, to protect database from hackers there are security techniques must implemented to avoid system from attackers.

4.1 Techniques to fight with SALIA

The detection approaches for SQLIA can be categorized broadly into pre-generated and post-generated approaches. Post-generated approaches are generally useful while analyzing dynamic SQL which is generated by web application such as Positive tainting and Syntax aware evaluation, Context Sensitive String Evaluation, Parse tree

evaluation based on grammar and DUD [Debasish, Utpal and D.K. Bhattacharya] approach. Pre-generated approaches are generally used during the testing phase of the web application such as Pixy and Program Query Language. Also Application layer intrusion detection approach which breaks data into buckets as done in network intrusion detection system. Relative frequencies of those buckets are used to compare with the historical data to decide about the intrusion, or Use prepared statements and parameterized queries to fight with SQLIA [3 & 10].

4.2 Digital Certificate and PKI

Digital certificate are electronic files that are used to identify people and resources over networks such as the Internet. Digital certificates also enable secure, confidential communication between two parties using encryption. Public Key Infrastructure (PKI) provides the core framework for a wide variety of components, applications, policies and practices to combine and achieve the three principal security functions (integrity, authentication and nonrepudiation). A PKI is a combination of hardware and software products, policies and procedures. It provides the basic security required for secure communications so that users who do not know each other or are widely distributed, can communicate securely through a chain of trust. Digital certificates are a vital component in the PKI infrastructure as they act as 'digital passports' by binding the user's digital signature to their public key [26].

4.3 Encryption

Encryption / It prevents exposure of sensitive information even if the database server is compromised so that when a database is compromised by an intruder, data remains protected even when a database is successfully attacked or stolen. Furthermore, database encryption can be employed to maintain the data integrity, ensuring that even a little modification made on the data can be detected. Database encryption technology meets the data confidentiality requirements and has become an indispensable aspect of enterprise database security [27].

4.4 Http Proxy Server Firewall

The user contacts the gateway using a TCP/IP application, such as Telnet or FTP, and the gateway asks the user for the name of the remote host to be accessed. When the user responds and provides a valid user ID and authentication information, the gateway contacts the application on the remote host and relays TCP segments containing the application data between the two endpoints. If the gateway does not implement the proxy code for a specific application, the service is not supported and cannot be forwarded across the firewall [28].

4.5 Access Control Mechanism

Access Control Mechanism is a technique to maintain data confidentiality. When someone tries to access data object, Access Control Mechanism checks the rights of the user against set of authorizations. They are generally specified by security administrator or security officer. Authorizations are given as per the security policy of the organization. Along with Access Control Mechanism, A strong Authentication mechanism is also required to authenticate the valid user of a database system. After that access control will help defining different access permissions on different data objects of a database [29].

4.6 Enforcing Buffer Size Limitation

An effective way to prevent an overflow is to strictly enforce the buffer's size limitation. Simply stated, never allow more data to be placed into a buffer than it is designed to hold. Stack validation, a critical part of an overflow attack is modifying the return address pushed onto the stack by the caller. Once the called procedure returns using the altered return address, control is passed to the attacker's code and the attack succeeds. If the called procedure could detect the stack tampering, the application could terminate itself before executing the attacker's code. By pushing a static value onto the stack and validating it before returning, a called procedure can avoid passing control to malicious code. These static values are often called *static canaries* or *canary values*, and are used in products such as StackGuard and the Immunix Secured OS. When the buffer is overflowed to change the return address, the canary value is overwritten because it is located between the buffer and the return address. By checking the value of the canary before returning from the procedure, it is possible to thwart the attack by terminating the process before the attacker's code is executed [30].

4.7 Data Scanning and Analyzing Tools

Security managers have turned to scanning and analysis tools to identify a wide variety of potential problems on their networks. While host-oriented patch tools such as Update EXPERT from St. Bernard Software and HFNetChkPro from Shavlik Technologies focus on the myriad patches needed to keep Windows servers up to date, network vulnerability analyzers look for more than just missing patches. These tools can search for misconfigured application servers, such as Web servers; and network components, such as switches and routers. They look for out-of-date applications, especially those

with known problems. And they often search for applications that are enabled by default--but perhaps shouldn't be, such as RPC services on UNIX or the UDP ECHO program on Windows NT/2000. They often look for "information leakage" from systems through DNS and other avenues, including SNMP and Windows registry [31].

4.8 Discretionary Access Control (DAC), Mandatory Access Control (MAC), and Role-Bases Access Control (RBAC)

Access control mechanisms of current DBMSs are based on discretionary policies governing the accesses of a subject to data based on the subject's identity and authorization rules. Mandatory access control (MAC) policies regulate accesses to data by subjects on the basis of predefined classifications of subjects and objects in the system. Objects are the passive entities storing information, such as relations, tuples in a relation, or elements of a tuple. Subjects are active entities performing data accesses. RBAC models represent arguably the most important recent innovation in access control models. RBAC models are based on the notion of role. A role represents a specific function within an organization and can be seen as a set of actions or responsibilities associated with this function. Under an RBAC model, all authorizations are granted to the role associated with that activity, rather than being granted directly to users. Users are then made members of roles, thereby acquiring the roles' authorizations. User access to objects is mediated by roles; each user is authorized to play certain roles and, on the basis of the roles, he can perform accesses to the objects [29].

4.9 Anti-Phishing Software

Over the past few years we have seen an increase in "semantic attacks" — computer security attacks that exploit human vulnerabilities rather than software vulnerabilities. Phishing is a type of semantic attack in which victims are sent emails that deceive them into providing account numbers, passwords, or other personal information to an attacker. Typical phishing emails falsely claim to be from a reputable business where victims might have an account. Victims are directed to a spoofed web site where they enter information such as credit card numbers or Social Security Numbers. Billions of dollars are lost each year due to unsuspecting users entering personal information into fraudulent web sites. To respond to this threat, software vendors and companies with a vested interest in preventing phishing attacks have released a variety of "antiphishing tools." For example, eBay offers a free tool that can positively identify the eBay site, and Google offers a free tool aimed at identifying any fraudulent site. As of September 2006, the free software download site **Download.com**, listed 84 anti-phishing tools [32].

4.10 Anti-Malware Tools

Anti-malware refers to software tools and programs designed to identify and prevent malicious software, or malware, from infecting computer systems or electronic devices. Anti-malware tools can also include malware removal capabilities, and the term anti-malware can range from code integrated with other software programs or in the operating system itself to third-party tools that scan for and remove a wide variety of malware variants. Also anti-malware software is commonly thought of as software tools for desktops and laptops, but anti-malware tools also abound for servers, workstations and mobile devices like smartphones and tablets [33].

4.11 Rootkit Detector and Remover tools

Sophos Virus Removal Tool will scan your computer and let you safely and reliably detect and remove any rootkit that might have hidden itself on your system. As part of its complete protection of endpoint computers, Sophos End user Protection has an integrated detection functionality that removes and prevents them being installed onto your desktops, laptops and servers [34].

4.12 Intrusion Detection System

Intrusion detection is a security technology that attempts to identify either individual who is trying to break into system and misuse information without authorization and\or those who have legitimate access to the resource but are taking undue advantage of their rights. The job of Intrusion Detection System (IDS) is to dynamically monitor the events occurring in a system and alert when any suspicious activity occurs so that defensive action can be taken to prevent or minimize damage. In general, the main goal of IDS is to detect malicious transactions before they are being committed and then dropping and rolling them back. Intrusion detection systems serve three essential security functions: they **monitor**, **detect** and **respond** to unauthorized activity [35].

4.13 File Integrity Monitoring

File Integrity Monitoring (FIM) is an internal control or process that performs the act of validating the integrity of operating system and application software files using a verification method between the current file state and the known, good baseline. This comparison method often

involves calculating a known cryptographic checksum of the file's original baseline and comparing with the calculated checksum of the current state of the file. The Verisys File Integrity Monitoring system provides a simple solution to your integrity monitoring requirements, giving you confidence that the integrity of your data has not been compromised [36].

4.14 Database Activity Monitoring

Database Activity Monitoring (DAM) is a database security technology for monitoring and analyzing database activity that operates independently of the database management system (DBMS) and does not rely on any form of native (DBMS-resident) auditing or native logs such as trace or transaction logs. DAM provides privileged user and application access monitoring that is independent of native database logging and audit functions. It can function as a compensating control for privileged user separation-of-duties issues by monitoring administrator activity. DAM is a powerful solution that independently monitors and audits all database activity across multiple database platforms. It provides an easy-to-use audit trail policy for all sensitive tables and columns, administrative access, and a "before and after" view of all changes. Some DAM solutions include full monitoring of applications and other sources of database calls [37].

4.15. Database Firewall

A firewall forms a barrier through which the traffic going in each direction must pass. A firewall security policy dictates which traffic is authorized to pass in each direction. A firewall may be designed to operate as a filter at the level of IP packets, or may operate at a higher protocol layer. Firewalls can be an effective means of protecting a local system or network of systems from network-based security threats while at the same time affording access to the outside world via wide area networks and the Internet. A firewall may act as a packet filter. It can operate as a positive filter, allowing passing only packets that meet specific criteria, or as a negative filter, rejecting any packet that meets certain criteria. Depending on the type of firewall, it may examine one or more protocol headers in each packet, the payload of each packet, or the pattern generated by a sequence of packets [28].

4.16 SSL and WTLS

Secure Sockets Layer (SSL) technology is a security protocol that is today's de-facto standard for securing communications and transactions across the Internet. SSL has been implemented in all major browsers and Web servers, and as such, plays a major role in today's e-commerce and e-business activities on the Web [38]. The Wireless Application Protocol (WAP) is a standard to provide mobile users of wireless phones and other wireless terminals access to telephony and information services, including the Internet and the Web. WAP security is primarily provided by the Wireless Transport Layer Security (WTLS), which provides security services between the mobile device and the WAP gateway to the Internet [28].

5. Allocation Proposed Security Techniques

Proposed resource allocation approach first: determined which security techniques to assign to each database threat based on experimental analysis, this analysis based on features of database threats from perspective of hackers. In this research are made experimental analysis on 22 previous threats and vulnerabilities, this analysis was based on some features/characteristics of database threats such as how attacker made unauthorized access on database system at each threat, Attack type, Location of threats, Exploitability, Impact, Users, and Scope. Attack type feature refer to Active or Passive attack, all threats are active attack but inference threat is only passive attack, and when this attack happen which attack in database system (Database Server, Web Server, Browser, Network Infrastructure, and Operating System Attack). Location of threats feature refer to where this attack may happen such as Internal mean that this attack happened inside the system (Intra-organization- Local Area Network (LAN)) and External mean that this attack happened outside the system (Extra-organization-WAN Network). Exploitability feature refers to ability users to hack system such as Easy, Average, Difficult, and Very Difficult. Impact feature refer to degree effect threat on system such as Severe and Moderate, in this research not focused on low threats impact as these threats not important. Users feature refer to who can made unauthorized access on system, authorized means that attacker privileged but made misuse or abuse data and unauthorized means that attacker unprivileged access system as privileged user. Scope feature refer to impact threat on security services such as confidentiality, access control, integrity, and availability. All these features of threats are shown in previous experimental study.

This allocation approach **second**: decide where objects (Security techniques) are allocated free for database threats based on location of threat and impact of threat on system. In this research threats classified into internal threats and external threats based on source of threats. In this research database threats classified into internal database threats and external database threats, Table (2) determines internal

threats which happen inside organization, and proposed security techniques which must implemented on data inside system such as Use prepared statements and parameterized queries (SQLIA), Anti-malware tools, Rootkit Detector and Remover tools, Data and Memory Encryption, Data scanning and analyzing tools, and Access Control or on internal communication of LAN Network such as all other proposed security techniques in table (2), also these threats arranged according to impact threats.

Table 2: Internal Database Threats

Threats	Impact	Proposed Security Techniques
SQL Injection	Severe	Use prepared statements and parameterized queries, and use Pre-generated approaches, Post generated approaches, Application layer intrusion detection approach, and SAFELI approach
Weak Authentication	Severe	Use Certificates and PKI
Unmanaged Sensitive Data	Severe	Data Encryption Access control
Back up Data Exposure	Severe	Data Encryption
Advanced Persistent Threat	Severe	Deploy Memory/Data Injection Prevention Technologies Memory and Network Encryption
Malware	Severe	Anti-malware tools
Database Rootkits	Severe	Rootkit Detector and Remover tools
Excessive Privilege Abuse	Severe	Deploying IDS to detect Insider Attacks
Legitimate Privilege Abuse	Severe	Deploying IDS to detect Insider Attacks
Privilege Elevation	Severe	File integrity monitoring
Database Platform Vulnerabilities	Severe	Database Activity Monitoring Database Firewalls
Database Communications Protocol Vulnerabilities	Severe	Use SSL & WTSL
Unpatched DBMS	Moderate	Data scanning and analyzing tools
Redundant DBMS Features Enabled	Moderate	Data scanning and analyzing tools
Broken Configuration Management	Moderate	Data scanning and analyzing tools
Inference	Moderate	MAC, DAC, and RBAC
Weak Database Audit Trial	Moderate	Audit duties should ideally be separate from both database administrators and the database server platform to ensure strong separation of duties policies

Table (3) determines external threats which happen outside organization, and proposed security techniques which must implemented to protect system, these techniques implemented on user's system such as Use prepared statements and parameterized queries (SQLIA), Anti-malware tools, Rootkit Detector and Remover tools, and Use anti-phishing software or implemented on external communication of internet such as all other techniques, also these threats arranged according to impact threats.

Table 3: External Database Threats

Threats	Impact	Proposed Security Techniques
SQL Injection	Severe	Use prepared statements and parameterized queries, and use Pre-generated approaches, Post generated approaches, Application layer intrusion detection approach, and SAFELI approach
Weak Authentication	Severe	Use Certificates and PKI
Buffer Overflow Attack	Severe	Strictly enforce the buffer's size limitation
Advanced Persistent Threat (APT)	Severe	Deploy Memory/Data Injection Prevention Technologies Memory and Network Encryption
Malware	Severe	Anti-malware tools
Database Rootkits	Severe	Rootkit Detector and Remover tools
Database Communications Protocol Vulnerabilities	Severe	Use SSL & WTSL
Web Application Attacks	Moderate	Use http proxy servers firewalls
Covert Channel	Moderate	Use resource monitoring techniques
Inference	Moderate	MAC, DAC, and RBAC
Social Engineering	Moderate	Use anti-phishing software

From two previous tables, there are threats which occurred inside and outside organization such as SQL Injection, Weak Authentication, Advanced Persistent Threat, Inference, Malware, Database Rootkits, and Database Communications Protocol Vulnerabilities. Also Denial of Service attack doesn't have supported techniques as it is difficult to prevent DOS attack but can discover the reason of this attack and solve the problem.

6. Conclusions

As databases hold a critical concentration of sensitive information, and as a result, databases are vulnerable, so database systems become the favorite target for hackers. In this paper vulnerabilities and threats which may face database system are survived and from previous analysis concluded that these threats impact on database system for all in database server, web server, browser, network infrastructure, and operating system, these mean that all part of database system become attacked from hackers. So today, enhancing the security of database is becoming one of the most urgent tasks in database research and industry to protect database system and to prevent compromised database. Also Audit duties should ideally be separate from both database administrators and the database server platform to ensure strong separation of duties policies (Regulatory Problem). Also must use resource monitoring which obtaining information concerning the utilization of one or more system resources and it is used to monitor the change in computer resources that caused by malware execution**.**

References

[1] Sabrina. D. C. V, Pierangela. S, Sushil. J, 1999, "Database Security", "European Community within the FASTER Project in the Fifth (EC) Framework Programme under contract IST-1999-11791", pp. 1-21

[2] Nedhal A. Al and Dana. Al, 2013, "Database Security Threats: A Survey Study", "International Conference on Computer Science and Information Technology (CSIT)", pp.60-64

[3] Saurabh. K and Siddhaling. U, 2012, "Review of Attacks on Databases and Database Security Techniques", "International Journal of Emerging Technology and Advanced Engineering", pp.253-263

[4] Erez. S, Ronen. V, Ehud. G and Yuval. E, 2014," Implementing a database encryption solution, design and implementation issues", "computers & security", pp. 33 – 50

[5] Ponemon Institute, 2014, "2014 Cost of Data Breach Study: Global Analysis / Research Report", "IBM - Ponemon Institute LLC", pp.1-28

[6] Imperva's Application Defense Center, 2013, "Top Ten Database Security Threats" "Data Security for the Data Center", pp.1-11

[7] Imperva's Application Defense Center, 2014, "Top Ten Database Security Threats" "Data Security for the Data Center", pp.1-9

[8] OWASP, 2013, "The Ten Most Critical Web Application Security Risks", "the Open Web Application Security Project - OWASP", pp. 1-22

[9] Amichai. Sh, 2006, "Top Ten Database Security Threats", "CTO Imperva, Inc.", pp.1-14

[10] Shivnandan. S and Rakesh. K. R, 2014, "A Review Report on Security Threats on Database", "(IJCSIT) International Journal of Computer Science and Information Technologies, pp. 3215 – 3219

[11] Abdul Razzaq, Ali. H, Nasir. H and Farooq. A, 2009, "Multi-Layered Defense against Web Application Attacks", "Sixth International Conference on Information Technology: New Generations", pp.492-497

[12] Andrew. B, Dan. B and Palash. N, 2007, "Exposing Private Information by Timing Web Applications", "the International World Wide Web Conference Committee (IW3C2)", pp.1-8

[13] Crispin. C, Perry. W, Calton. P, Steve. B and Jonathan. W, 1999, " Buffer Overflows: Attacks and Defenses for the Vulnerability of the Decade", "IEEE, and Proceedings of DARPA Information Survivability Conference and Expo (DISCEX)", pp.1-11

[14] James. C. F, Vitaly. O, Nish. B and Niels. H, 2005, " Buffer Overflow Attacks: Detect, Exploit, Prevent ", "Syngress, Inc.", pp.1-521

[15] Damballa, Inc., 2010, "Advanced Persistent Threats (APTs)", available at "https://www.damballa.com/advanced-persistent-threats-a-brief-description/" 9/1/2015

[16] Sam. M, 2014, "Advanced Persistent Threat – APT", available at "https://www.academia.edu/6309905/Advanced_Persistent_Threat_-_APT" 9/1/2015

[17] Steven. J. M, 2007, "Technical Report: Covert channel vulnerabilities in anonymity systems", "UCAM-CL-TR-706 / ISSN 1476-2986", pp.1-140

[18] Mark. T, 2012, "Top 10 Database Vulnerabilities and Misconfigurations", "APPLICATION SECURITY, Inc.", available at "http://www.sifma.org/uploadedfiles/societies/sifma_internal_auditors_society/top10-database-vulnerabilities-and-misconfigurations.pdf" 1/1/2014

[19] Salvador. M, 2000, " Inference Attacks to Statistical Databases: Data Suppression, Concealing Controls and Other Security Trends", Aleph Zero online magazine, number 23", pp.1-12

[20] Emil. B, 2009, "DATABASE SECURITY - ATTACKS AND CONTROL METHODS", "JAQM: JOURNAL OF APPLIED QUANTITATIVE METHODS / Software Analysis", pp.499-454

[21] Lech. J. J and Lingyan.R. F, 2010, "Social Engineering-Based Attacks: Model and New Zealand Perspective", "IEEE, International Muliconference on Computer Science and Information Technology", pp.847-853

[22] Jim. R, 2012, "The Ongoing Malware Threat: How Malware Infects Websites and Harms Businesses — and

What You Can Do to Stop It", "Symantec Corporation - VeriSign, Inc, pp.1-11
[23] John. H, 2006, "Rootkit threats", "NGS – New Generation Software", pp.18-19
[24] Iqra. B, Farooque. A, and Abdul Wahab. M, 2012, "Database Security and Encryption: A Survey Study ","International Journal of Computer Applications (0975 – 888)", pp.28-34
[25] Kevin. K, 2006, "Cryptography in the Database: The last line of Defense", "USA, Symantec Corporation", pp.4-11
[26] Ray. H, 2001, "PKI and Digital Certification Infrastructure", "9th IEEE International Conference on Networks (ICON.01)", pp. 234 – 239
[27] Gang. Ch, Ke. Ch, and Jinxiang. D, 2006, "A Database Encryption Scheme for Enhanced Security and Easy Sharing", "10th International Conference on Computer Supported Cooperative Work in Design", pp. 1-6
[28] William. S, (2011), "Cryptography and Network Security Principles and Practices, Fifth Edition", "publishing as Prentice Hall", pp. 1-900
[29] Elisa. B, and Ravi. S, 2005, "Database Security—Concepts, Approaches, and Challenges ", "IEEE Transactions on Dependable and Secure Computing", pp. 2-19
[30] Jason. D, 2004, "Defeating Overflow Attacks", "SANS Institute InfoSec Reading Room", pp. 1-30
[31] Joel. S, 2003, "Testing and comparing vulnerability analysis tools", "TechTarget", available at "http://searchsecurity.techtarget.com/Testing-and-comparing-vulnerability-analysis-tools" 2/5/2015
[32] Yue. Z, Serge. E, Lorrie. C and Jason. H, 2006, " Phinding Phish: Evaluating Anti-Phishing Tools", " Carnegie Mellon University / Human-Computer Interaction Institute by an authorized administrator of Research Showcase ", pp. 1-17
[33] Forrest. S, 2015, "Anti-Malware", "webopedia", available at "http://www.webopedia.com/TERM/A/anti-malware.html" 2/5/2015
[34] Sophos Ltd., "2015", available at https://www.sophos.com/en-us/products/free-tools/virus-removal-tool.aspx, 1/6/2015
[35] Alka. J and Sweta. J, 2010, "Database Intrusion Prevention cum Detection System with Appropriate Response", "International Journal of Information Technology and Knowledge Management", pp. 651-656
[36] Ionx Solutions LLP, 2015, "Verisys product", available at http://www.ionx.co.uk/solutions/file-integrity-monitoring 30/4/2015
[37] Mark. N, Avivah. L and Paul. E. P, 2009, "Pattern Discovery with Security Monitoring and Fraud Detection Technologies", "Gartner Inc.", pp. 1-10
[38] Entrust, Inc., 2007, "Understanding Digital Certificates & Secure Sockets Layer: A Fundamental Requirement for Internet Transactions", "Entrust, Securing Digital Identities & Information", pp. 1-11

Prof. Abdel Nasser H. Zaied, is prof. of Information Systems, Dean, Faculty of Computers and Informatics, Zagazig University, Egypt. He previously worked as an associate professor of Industrial Engineering, Zagazig University Egypt, an assistant professor of Technology Management, Arabian Gulf University, Bahrain; and as visiting professor at Oakland University, USA. He supervised 12 PhD. thesis and 45 MSc. thesis, and examined 8 PhD. thesis and 47 MSc thesis. He published 30 research papers in International and Regional Journals and 22 research papers in International and National conferences. His areas of research are: Systems Analysis and Design; Information Security; Knowledge Management; Quality Management Systems, Information Security and project Management, Electronic applications.

Prof. Walid I. Khedr is an associate professor of Information Technology, Head of Information Technology department, Faculty of Computers and Informatics, Zagazig University, Egypt. His current research interests are primarily in network security protocols, cryptography, key management protocols, and RFID security. Another field of interest is quantum cryptography.

Shimaa S. Mohamed is a Lecturer of Decision Support Systems and MSc. candidate, Faculty of Computers and Informatics, Zagazig University, Egypt.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Detecting Communities and Surveying the Most Influence of Online Users

Thanh Tran[1], Thanh Ho[2] and Phuc Do[1]

[1]University of Information Technology, VNU-HCM, Vietnam
duythanhcse@gmail.com, phucdo@uit.edu.vn

[2]Faculty of Information System, University of Economics and Law
VNU-HCM, Vietnam
thanhht@uel.edu.vn

Abstract

Social network is a virtual environment that provides services for connecting users with the same interests, points of view, gender, space and time. Beside connection, information exchange, communication, entertainment and so on. Social network is also an environment for users who work in online business, advertisement or politics, criminal investigation. How to know what users discuss topics via exchanged contents and communities which users join in? In this paper, we propose a model by using topic model combined with K-means to detect communities of online users. Each user in social network is represented by a vector in which the components are the distribution probabilities of interested topics of that user. Based on the components of this vector, we discover the interested topics of online users to detect communities and survey users who are the most influence in communities to recommend for spreading information on social network.

Keywords: *LDA, ART, K-Means, online community, topic model, influence.*

1. Introduction

The social network has become a familiar concept of information technology. Positive or negative impacts are shown through the analysis of social networks, which is much more important than the work of capturing information and settling information in the real social life. Actually, at present there are a lot of issued researches on the social network analysis [1][2][3][4].

The social structure of social networks represented like a human society in the real life is known as the online community [4][5][6]. A social network is a heterogeneous huge data set, with many links represented by a graph. In the graph, with the actor corresponding to the object and the edge corresponding to the link in the interactive relationship between the objects, a social network will have the similarities in online communities, such as the similarity in friend relationships, the similarity in interests, and the similarity in affinities of other characteristics, including work, education, social interaction [1][2][3]. The social network for clustering is to find out the characteristics similar to online communities put into groups according to specific topics.

Figure 1. Communities on social networks[1]

Clustering social networks has many implications in management, economic activities, science and society. Social networks with lots of data to analyze at present have three main types of data which are often analyzed as follows:

Firstly, the analysis is based on the friend relationship: this analysis, which is mainly based on the relationship of friends in community on the social network, has important implications in identifying the strength in community relations [5][6][7]. This helps managers make decisions effectively in their organizations and we can also determine the characteristics of a community when we know a few online communities like the expression "please tell me who is your friend, I will tell you how you are ".

Secondly, the analysis is based on the exchanged contents [7]. This analysis helps understand that the interest swap of users happens on social networks. Clustering the relevant exchanged contents really helps us cluster

[1] http://treeintelligence.com/en/influence-and-viralization-networks/

communities sharing their same opinions to strengthen collaboration.

Thirdly, clustering social network communities consists of structure and content [8]. This helps find out the online users in communities which have the same structure and content. From that, managers can easily communicate and create the most efficient group collaboration.

Based on the features of communities, we can find out communities users who are the most influence in communities.

Figure 2. Influential factor of community members[2]

In the paper, we propose a model for detecting interested topics via exchanged contents on the social network based on Latent Dirichlet Allocation model [9] and Author-Recipient-Topic [10], and then proceed detecting the community by algorithm K-Means [11] combined with the topic model.

The paper is organized as 1) Introduction 2) Related work 3) General model for detecting community based on K-means and topic moel 4) Experiments and discussions 5) Conclusion and future work.

2. Related work

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete text data [10]. In general, LDA is a three-level hierarchical Bayesian model [9][10][12] in which each document is described as a random mixture over a latent set of topics. Each topic is modeled as a discrete distribution of a set of words. LDA is suitable for the set of corpus and the set of grouped discrete data. LDA can be used for modeling the document on the purpose of detecting some underlying topics of that document. The generative process of a set of documents consists of three steps: (i) each document has a probabilistic distribution of its topics; this distribution is estimated as the Dirichlet distribution. (ii) for each word in a document, a specific topic based on the distribution of the topics of that document is chosen (iii) each keyword will be chosen from the multinomial distribution of the keywords according to the chosen topic [10][12].

The purpose of LDA is to detect each word belonging to a specific topic. From that we can guess the label of that topic. The importance of topic model is the posterior distribution. This can be seen as the generative process and the posterior inference for the latent set of variables, which are the keywords of the topic. In LDA, this process is calculated by the equation:

$$p(\theta, \emptyset, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \emptyset, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \quad (1)$$

In the equation (1), we have the variables z, θ, ϕ. For each θ_j which is a vector of topics of document j, z_i is the topic of word w_i, $\phi^{(k)}$ is the matrix KxV with $\phi_{i,j} = p(w_i \mid z_j)$

However, in equation (1), we can't precisely calculate with the normal factor $p(w \mid \alpha, \beta)$. Therefore, we normally use Gibbs Sampling (Griffiths & Steyvers, 2004; Steyvers et al., 2004; Rosen-Z vi et al., 2004) for inference.

2.2 Gibbs Sampling

Gibbs Sampling is a member of a family of algorithms from the Markov Chain Monte Carlo (MCMC) [13]. The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations through the chain, sampling from the distribution should converge to be close to sampling from the desired posterior. Gibbs Sampling [10][12][13][14] is based on sampling from conditional distributions of the variables of the posterior. For example, to sample x from the joint distribution $p(x) = p(x_1,x_2,\ldots,x_m)$. We do not have any proper solution to compute p(x), but a representation for the conditional distribution is possible, using Gibbs Sampling perform the following steps [12]:

1. Randomly initialize each x_i
2. For t=1…T:
   2.1. $x_1^{t+1} \sim p(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_m^{(t)})$
   2.2. $x_2^{t+1} \sim p(x_2|x_1^{(t)}, x_3^{(t)}, \ldots, x_m^{(t)})$
   2.3. $x_m^{t+1} \sim p(x_m|x_1^{(t)}, x_2^{(t)}, \ldots, x_{m-1}^{(t)})$

2.3 ART model (Author - Recipient - Topic)

ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent [5][10]. In its generative process for each message, an author a_d and a set

[2] http://treeintelligence.com/en/influence-and-viralization-networks/

of recipients $\boldsymbol{r}_d$ will be observed. To generate each word, a recipient x is chosen from the set $\boldsymbol{r}_d$; and then, a topic z is chosen from a multinomial distribution $\theta_{a_d,x}$. This distribution is specified with the *author-recipient* pair (a_d, x). Finally, the word w is generated by choosing samples from a multinomial distribution $\emptyset_z$. The process of choosing samples is based on the *Gibbs sampling* algorithm. The final result is the discovery of topics in a social network where the messages are created.

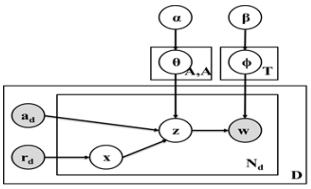


Figure 3. ART model [10]

In ART, given hyperparameters α and β, the author α_d and the set of recipients r_d,the joint distribution of an author mixture θ , a topic mixture $\emptyset$, a set of recipients x_d (belonging to X_d), a set of topics z_d (belonging to N_d), and a set of words w_d (belonging toN_d) is given by:

$$p(\theta, \emptyset, \boldsymbol{x}_d, \boldsymbol{z}_d, \boldsymbol{w}_d | \alpha, \beta, a_d, \boldsymbol{r}_d) \quad (2)$$
$$= p(\theta|\alpha)p(\emptyset|\beta)\prod_{n=1}^{N_d} p(x_{dn}|\boldsymbol{r}_d)p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\emptyset_{z_{dn}})$$

Integrating over θ and $\emptyset$, summing over $\boldsymbol{x}_d$ and $\boldsymbol{z}_d$, we get the marginal distribution of a document:

$$p(\boldsymbol{w}_d|\alpha, \beta, a_d, \boldsymbol{r}_d) \quad (3)$$
$$p(\theta|\alpha)p(\emptyset|\beta)$$
$$= \iint \prod_{n=1}^{N_d} \sum_{x_{dn}} \sum_{z_{dn}} p(x_{dn}|\boldsymbol{r}_d)p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\emptyset_{z_{dn}})d\emptyset d\theta$$

Finally, we have the probability of a corpus is:

$$p(D|\alpha, \beta, a, r) = \prod_{d=1}^{D} p(w_d|\alpha, \beta, a_d, r_d) \quad (4)$$

ART model describes the interaction of each node by analyzing transferring information of each node in the network; a topic relates to author, recipient and discovers role of author and recipient in transferring information process. Hence, the identification of topics in ART model depends on the social network in which messages are sent and received. Each pair of sender and receiver has a distribution over topics and each topic has a distribution over words.

2.4 Community-Author-Recipient-Topic model (CART)

In [15], the authors introduce CART model (Community - Author - Recipient - Topic), the model is tested on the Enron email data system [3]. The model shows that the discussion and exchange between users within a community are related to the other users in community. This model is binding on all relevant users and the topics discussed in the emails belonging to a community, while the same users and the various topics can link to other communities. Compared with the above models including CUT, CART model is closer to further emphasize the ways that the topics and their relationships affect the structure of the online community in exploring community on topics [15].

2.5 Finding the cluster of actors model

In [16], the authors present how to use SOM network to cluster the actor based on interested topic vector from dataset in English. This vector is a distribution probability of topic that actor prefers. The authors use ART model to create the vector of interested topics and use Enron email corpus as a sample dataset to evaluate efficiency in SOM network. By experimenting on the dataset, the authors demonstrate that our proposed model can be used to extract well and meaningful cluster following the topics [16].

2.6 Motivation research

We propose a model by using K-means algorithm combined with topic model for clustering online users based on their interested topics to detect online communities. These topics are exploited from a corpus of messages in Vietnamese on social networks via exchanged contents of users. Besides that, we survey the users who are the most influence in communities for spreading information on social network.

3. General model

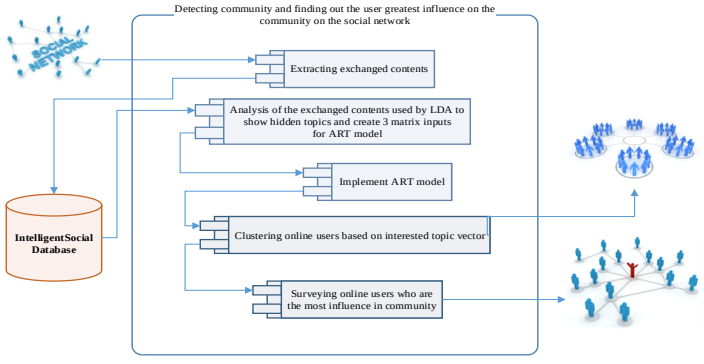


Figure 4: Model of detecting communities and finding out the users with the most influence in communities

[3] https://www.cs.cmu.edu/~./enron/

We propose a model as shown in Figure 4. Our model consists of information extraction, data cleaning, social network analysis to find out the interested topics by using the LDA model, labeling topic and detecting online community. There are 5 steps, including:

Step 1: We do the data cleaning process for the social networks dataset. Each message will be characterized by keywords and removed the stop words.

Step 2: After cleaning the data and using the LDA model, we will have the matrix words of the topics T x V (word, the distribution probability) and the matrix distributed the messages based on the topics T x D (message's id, the distribution probability) .

Step 3: Applying the matrixes T x V and T x D for ART model to create interested topic vectors. Each online user has a interested topic vector. Each vector of online users based on interested topics is a vector representing the interested probability of the topics of each user in a social network. Each user can have one or many interested topics.

Step 4: Using K-means algorithm for clustering of users on social networking based on interested topic vector created in step 3.

Step 5: Surveying users who are the most influence in communities to recommend for spreading information on social network.

4. Experiment and discussion

4.1 Input dataset

The dataset is collected from Facebook, include:

- 75740 posts and comments in Vietnamese
- 2315 online users.
- 10 topics are surveyed.
- From 2014 – 2015 (2 years).

4.2 Implementation

After cleaning the data, by using the LDA model with the parameters $\alpha = 0.5$, $\beta = 0.1$, the number of iterations for Gibbs sampling is 2000, the number of steps is 100, the number of topics is 50 [10][12][14]. We have the matrix words of the topics T x V (word, the distribution probability).

After using the LDA model to figure out the topics, we use K-means algorithm [11] to group the messages into cluster. We consider each cluster as community. There are *n* users, each user has *m* attributes, and we divide them into *k* communities based on their attributes by using the K-means algorithm.

For clustering exchanged contents: we analyze the exchanged contents of users on social networks to find out interested topics vector, and then use the K-Means algorithm to cluster community based-on interested topics vector of users. In a community includes users who have the same interested topics. To do this, we study and implement the model ART (author – recipient – topic).

During the analyzing process, the ART model will create three matrices: the distribution matrix of words according to the distribution matrix of topics, authors, recipients and messages.

a. Matrix 1: The message - author – recipient (table 1)

This matrix contains message - author - recipient, each line consists of message code, author code and recipient code. To get this matrix, we extract and analyze information from tables of posts, comments and profile of users in the dataset.

Table 1: Matrix message - author - recipient

Message (ID)	Author (ID)	Recipient (ID)
629	200	196
630	200	196
631	200	196
632	222	196
633	219	196
634	222	196
635	222	196

b. Matrix 2: message - vocabulary - frequency of appearance

Table 2 – Matrix message - vocabulary - frequency of appearance

Message (ID)	Vocabulary (ID)	Frequency of appearance
633	220	1
633	20636	1
633	65	3
633	5343	1
634	15084	1
634	16273	1

Each line of this matrix contains the message ID, vocabulary ID and frequency of appearance in the message. This matrix usually contains volume of data very large because it must analyze each message contained in Matrix 1, for each message must parse out the matrix 2.

c. Matrix 3: Vocabulary – Vocabulary ID

Building more than 23,000 words extracted from the VnTokenizer tool, this is a project under the state was announced.

Table 3 - Matrix vocabulary – vocabulary ID

Vocabulary	Vocabulary (ID)
khử_trùng	5314
vội_vã	6038
Josu	6970
nhân_quyền	16488
bản_thân	14221
nhoè_nhẹt	8062
Hai_Hoàng	17589
Dương_Minh_Quang	6408
Tanimex	6971
Lê_Dũng	8265
ngập_chìm	17350

4.3 K-means algorithm combined with topic model for clustering communities

We use the K-means with the parameters: 10 topics, the number of messages is 75740, 2315 users, the number of topics is 10, the number of communities is 4 (k=4). We have experimented with k = 2, k = 3, k = 4, k = 5, k = 6,.., k = 10 ... to have a comparable clustering results and choose k = 4.

Figure 5 shows in detail the number of users (actors) belonging to the cluster (4 clusters):

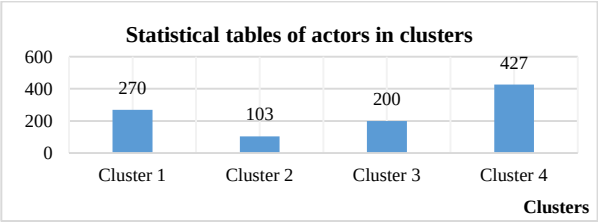


Figure 5 - The number of users in each cluster and the number of clusters. Cluster 1 has 270 users, cluster 2 has 103 user, cluster 3 has 200 users and cluster 4 has 427 users.

Table 4 – The set of vector centroid of 4 clusters

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
C1	0.00	0.00	0.00	0.28	0.01	0.02	0.02	0.43	0.00	0.24
C2	0.00	0.11	0.16	0.05	0.04	0.01	0.19	0.00	0.34	0.10
C3	0.49	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.20
C4	0.70	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

From the data in table 4, we easily know how much each cluster interested topic vectors and Distribution the topics according to vector centroid of 4 clusters (communities).

Figure 6 shows 4 communities (C1, C2, C3 and C4), each community has topics. For example, community C1 has 6 topics (T3, T4, T5, T6, T7 and T9), C2 has 8 topics (T1, T2, T3, T4, T5, T6, T8 and T9), C3 has 4 topics (T1, T2, T8 and T9) and C4 has 2 topics (T1 and T2). In which, C2 has the number of topics more than C1, C3 and C4.

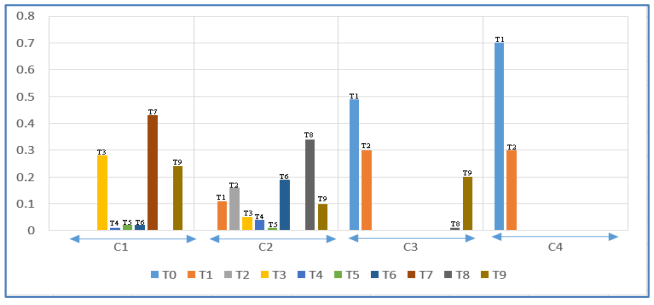


Figure 6 - Distribution 10 topics (from T1 to T9) according to vector centroid of 4 clusters (communities)

Figure 7 and figure 8 show result of detecting communities. There are 4 communities related to 4 clusters.

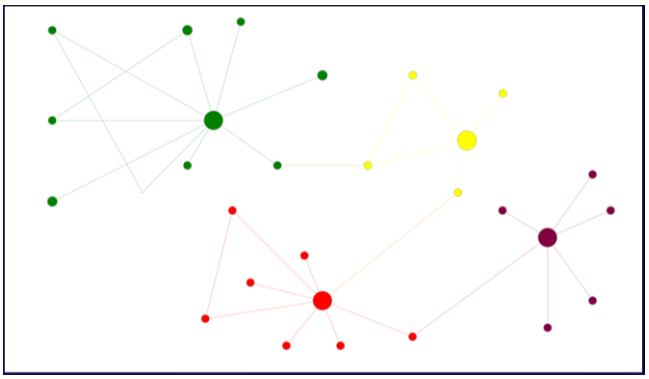


Figure 7. Results of detecting communities without user's name on nodes and surveying user who has the most influence in communities

.

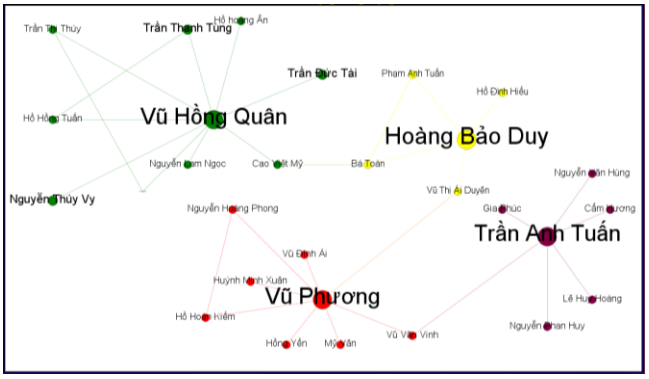


Figure 8. Results of detecting communities and surveying user who has the most influence in community

Each community was distinguished by a different color, if the user has a relationship of other users, they will be linked together. In the illustration above, we have 4 clusters with 4 users who are the most influence in communities such as: "Vũ Hồng Quân" in community 1, "Hoàng Bảo Duy" in community 2, "Vũ Phương" in community 3 and "Trần Anh Tuấn" in community 4 (see figure 8).

We can see the result {vuhong.quan.73, tough.crystal, nhung.vu.58760, ...} of cluster 1, {hoangbaoduy, transleyhan, nguyenhieu08 ...} of cluster 2 and so on (see table 5).

5. Conclusions and future work

5.1 Conclusions

Our research has proposed propagation models and algorithms through the analysis of social networks based on the specific topics. Especially, the research focuses on finding the topics with LDA model and clustering exchanged contents by using K-means algorithm combined with the topic model.

- Building the automated tools to retrieve exchanged contents from Facebook: All the exchanged contents consist of online users, posts, comments, likes, etc.
- Experimenting the detection topic module with LDA model, will help select a number of interested topics, such as political security, science and technology, sports, culture, arts, health and education to carry out clustering community on social networks.
- Experimenting Author - Recipients - Topics (ART) model will help create the set of vectors with interested topics of online users to provide community clustering.
- Proposing the community clustering model by using K-means algorithm combined with the topic model is to cluster online users based on interested topics vectors to find out online communities. After detecting community, we survey the user who has the most influence in community in order to recommend for spreading information on social networks.

5.2 Future work

In the future, we will continue to study and give recommendations in order to evaluate the results of the cluster as well as the quality of the proposed model.
Clustering the exchanged contents of online users helps us find out communities on social networks. In this community, there will be the same interested topics of online users who will have no relationship with other online users. Therefore, in order to cerate the relationship between the online users having the same interested topic with other online users, we will propose a model to spread the information to other online users on social network based on the users who are the most influence in community.

Acknowledgments

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-26-02.

Table 5 - Matrix probability of interest topic vectors of online users

Topic / Online users	T-0	T-1	T-2	T-3	T-4	T-5	T-6	T-7	T-8	T-9
vuhong.quan.73	0.00600	0.00110	0.0031	0.2000	0.05000	0.1000	0.1008	0.350	0.000	0.189
tough.crystal	0.00601	0.00105	0.0030	0.2020	0.05194	0.1000	0.0990	0.347	0.000	0.190
nhung.vu.58760	0.00700	0.00120	0.0030	0.2010	0.05000	0.1010	0.0868	0.350	0.000	0.200
motcoidive.hue	0.00150	0.00000	0.0000	0.2985	0.00000	0.0000	0.0000	0.450	0.000	0.250
hoangbaoduy	0.00000	0.00200	0.0000	0.0000	0.00100	0.0000	0.4970	0.000	0.300	0.200
transleyzhan	0.00000	0.00000	0.5801	0.0000	0.00000	0.0209	0.0000	0.000	0.399	0.000
nguyenhieu08	0.00000	0.00000	0.5799	0.0000	0.00000	0.0201	0.0000	0.000	0.400	0.000
vu.phuong.5264	0.50000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
nguyen.vietanh.338	0.51000	0.29000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
diem.dtk	0.50000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.200
Tuan777	0.70000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.000
truclieu.nguyen.3	0.70000	0.30000	0.0000	0.0000	0.00000	0.0000	0.0000	0.000	0.000	0.000

References

[1] J.Leskovec, L.A.Adamic, and B.A.Huberman, (2007). "The dynamics of viral marketing". In ACM Trans, volume 1.
[2] Muon Nguyen, Thanh Ho, Phuc Do (2013), *Social Networks Analysis Based on Topic Modeling*, The 10th IEEE RIVF International Conference on Computing and Communication Technologies (P.119-122), Hanoi.
[3] P.Domingos and M.Richardson, (2001), "Mining the network value of customers". In Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 01, pages 57–66, New York, NY, USA.
[4] Mr. Sachan, D. Contractor, T.A. Faruquie, L. V.Subramaniam, (2009), Using Content and Interactions for Discovering Communities in Social Networks. April 16–20, 2012, Lyon, France.
[5] Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang (2007), Topic and Role Discovery in Social Networks, Journal of Articial Intelligence Research 29.
[6] M. Sachan, D. Contractor, T. A. Faruquie, and L. V.Subramaniam. Probabilistic model for discovering topic based communities in social networks. 2011.

[7] Chunshan Li, William K. Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, Xin Li (2014), The Author-Topic-Community model for author interest profiling and community discovery, Springer-Verlag London 2014, pp. 74-85.
[8] The Anh Dang, Emmanuel Viennet (2012), Community Detection based on Structural and Attribute Similarities, ICDS 2012 : The Sixth International Conference on Digital Society, pp. 7-14.
[9] David M.Blei, (2003), "Latent Dirichlet Allocation". Computer Science Division, University of California, Berkeley, CA.
[10] Andrew McCallum, Andr´es Corrada, Xuerui Wang, (2004), The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email, Department of Computer Science, University of MA.
[11] http://en.wikipedia.org/wiki/K-means_clustering
[12] Tom Griffiths (2004), Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation - Gruffydd@psych.stanford.edu.
[13] B. Walsh, (2004), "Markov Chain Monte Carlo and Gibbs Sampling". *Lecture Notes for EEB 581, version 26 April 2004.*
[14] William M. Darling, (2011), "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling". *School of Computer Science University of Guelph.*
[15] N. Pathak, C. DeLong, A. Banerjee, and K. Erickson (2008), Social topic models for community extraction. The 2nd SNA-KDD Workshop, vol 8.
[16] Tran Quang Hoa, Vo Ho Tien Hung, Nguyen Le Hoang, Ho Trung Thanh, Do Phuc (2014), Finding the Cluster of Actors in Social Network based on the Topic of Messages, ACIIDS 04/2014, ThaiLan. Springer, pp. 183-190.

First Author. MS. Thanh Tran got master's degree at University of Information Technology, VNU-HCM, Vietnam. His strong ability is about website, mobile and desktop applications with the programming languages as Java, Php, Objective C, Android, and the database as MySQL and Oracle.

Second Author. MS. PhD Student. ThanhHo works for Faculty of Information System, University of Economics and Law, VNU-HCM, Vietnam. His interests are data mining, e-commerce, Business Intelligent, social network analysis and management information systems. He is a member of Prof. Do Phuc's project.

Third Author. Prof. Do Phuc works for the University of Information Technology, VNU-HCM, Vietnam. His interests are data mining, bioinformatics and social media analysis. His current project is toward the analysis of social network based on thecontent and structure.

ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18 , November 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Practical implementation of a methodology for digital images authentication using forensics techniques

Francisco Rodríguez-Santos[1], Guillermo Delgado-Gutierréz[1], Leonardo Palacios-Luengas[1] and Rubén Vazquéz-Medina[1,2]

[1] ESIME Culhuacan, Instituto Politécnico Nacional, Coyoacán, D. F. 04430, México
frodriguezs0901@alumno.ipn.mx

[2] CMP+L, Instituto Politécnico Nacional, Ticoman, D.F. 07340, México
ruvazquez@ipn.mx

Abstract

This work presents a forensics analysis methodology implemented to detect modifications in JPEG digital images by analyzing the image's metadata, thumbnail, camera traces and compression signatures. Best practices related with digital evidence and forensics analysis are considered to determine if the technical attributes and the qualities of an image are consistent with each other. This methodology is defined according to the recommendations of the Good Practice Guide for Computer-Based Electronic Evidence defined by Association of Chief Police Officers of UK; the methodology certainty level is verified by an efficiency coefficient, calculated by the quotient of the number of correct resolutions and the total number of analyzed images. This methodology can help to determine if a specific digital image can be used as evidence, and thereby, help to clarify events or incidents with legal, civil, administrative or criminal implications. Another advantage of the methodology is that it can be applied with open source software tools.

Keywords: *Forensic Science, Digital Evidence, Image Authenticity, Forensic Analysis Methodology, Digital Image Processing, Image Technical Attributes.*

1. Introduction

Today it is very common to find digital images due to the high availability of digital cameras in mobile phones. For some people, a picture may be irrelevant, but for some others, it may represent evidence which could be used to clarify facts with legal, civil, administrative or criminal implications. Therefore, a digital image could have a really high impact in our life and it could be much more representative than the oral or written description of an event, especially if it is considered that the description of that event could be distorted by a person, since time causes human memory deficiencies. With the technological advancement in mobile devices, the digital images have become ubiquitous today. However, modifying a digital image without any obvious traces is not a difficult task with the image editing software available these days. Grabler et al. [1] proposed a demonstration-based system for a visual step-by-step succinct generation tutorials of photo manipulations, which include changing the color of the eyes, teeth bleaching and enhancement of the sun setting, among others. Specialized software tools for digital images edition have potentiated the techniques of image manipulation. These tools allow almost everyone being able to improve the visual quality of an image in an effortless way according to their preferences, needs or interests. Also, these tools allow changing the perception of an event captured in a digital image. The motivations for these changes in digital images could be diverse. Some persons might edit a picture to have fun or to sell something. However, some others may try to involve someone in a wrongful act, or to obtain an illegal benefit.

Garry and Gerrie [2] showed that changing an image or improving its quality, may cause distortion of the reality perception, creating false records and affecting the memory of the people who watch it. Considering digital images that contain sensitive information that could be used as evidence, it is necessary to ensure the images' authenticity, in order to prevent that they are used in a malicious way to damage others. Farid and collaborators in different works showed techniques to determine if an image has been modified or not. Johnson and Farid in 2007 [3] described how such composites can be detected by estimating a camera's intrinsic parameters from the image of a person's eyes; Farid in 2009 [4] presented an overview of the passive techniques for detecting images forgery considering an image forensics context; Farid and Bravo in 2010 [5] showed that the visual system is remarkably inept at detecting simple geometric inconsistencies in shadows, reflections and perspective distortions, and they showed computational methods that can be applied to detect the inconsistencies that seem to elude the human visual system; Kee and Farid in 2010 [6] described a technique for measuring lighting conditions in an image, and described its use for detecting photographic composites; and finally, O´Brien and Farid in 2012 [7] described the existence of forensic techniques to detect geometric or statistical inconsistencies that result from specific forms of photo manipulation. Particularly, they

described a technique based on basic rules of image reflection and perspective projection.

There are several studies about image forensics methods that could help to determine the images' authenticity. Luo et al. [8] presented a survey and the implementation challenges about forensics passive technology. Hwang and Har in 2013 [9] proposed a re-interpolation algorithm which uses the characteristics of interpolation to detect forged images. Peng and Li in 2014 [10] proposed a method to identify among natural images, which represents a real fact, and which is a computer-generated graphics based on statistical and textural features. Hwang and Har in 2014 [11] showed that interpolation is an effective way to analyze digital images and define an identification method for digital image forgery and filtering region. On the other hand, Cao et al. [12] proposed an algorithm capable of concealing the quantization artifacts that are left in a single JPEG compressed image to hide the JPEG compression traces, which could make harder to find modifications in a digital image.

When an image is presented as evidence to clarify a sensitive case, it must be verified in order to determine if the fact that represents is real. Therefore, the process defined to verify the image authenticity must be robust, and it is based on international guides and best practices about evidence management.

This work proposes a methodology to determine if a JPEG image is authentic or not, and it is based on the features analysis of digital images using forensics techniques. The analyzed image features are the metadata, the image thumbnail, the camera traces derived from the demosaicing process and the signatures of software used to edit digital images. The demosaicing process allows reconstructing a full color image from the incomplete color samples output from an image sensor overlaid with a color filter array (CFA). The proposed methodology includes a set of methods that are applied independently, each one defines different evaluation metrics; which are used to define a technical resolution (verdict) that indicates if the analyzed digital image is authentic, post produced or modified. Finally, with this information, a technical dictum is generated in accordance to the NIST SP800-86 [13] guide.

2. Proposed methodology

The proposed methodology is a passive technique for image forensics that operates in the absence of any watermark or signature. The used techniques work on the assumption that although digital forgeries may leave no visual clues that indicate tampering, they may alter the underlying features of a digital image. The proposed methodology is not intended to detect specific changes in JPEG format images, it only determines if the analyzed images were modified or not, without specifying the used procedure and the image region that was modified. In addition, the performed analysis does not require the original image without any modifications or some extra attributes associated with it. The proposed methodology was basically defined according to the *Good Practice Guide for Computer-Based Electronic Evidence* defined by Association of Chief Police Officers of UK [14], although it also considers other international guidelines related with the digital forensic analysis and evidence management [15-20]. The proposed methodology considers that the analyzed images can be computer-based electronic evidence subjected to the same rules and laws that apply to documentary evidence.

The proposed methodology consists of 4 steps: 1) Collection, 2) Extraction of the image's technical features, 3) Analysis of the image's technical features, and 4) Issuance of the dictum.

2.1 Step 1: Collection.

This step includes two activities:

i) **Documentation.** The context of the incident, the device that generated the image, and the container device, in which the image is presented to be analyzed, must be documented. This is the first registration of the chain of custody on the methodology.
ii) **Saving and integrity verification.** The hash value (SHA 256) of the original image, and subsequently two identical copies of the original image must be obtained. It must be verified that the copies have the same hash value than the original image. One of these copies will be used for the analysis, and the other copy must be safely stored with the original image in order to support future comparisons. The registration of the validation that the three hash values are identical must be included in the custody chain of the process.

2.2 Step 2: Extraction of the image's technical features.

This step involves three activities:

i) **Format Verification.** The image format must be verified. It must be confirmed that the header of the image corresponds to a JPEG format. If the header does not match, the analysis process must be concluded and a register of this condition must be specified in the chain of custody of the process.
ii) **Features extraction.** If the image format matches the kind of image, then, the following information from the digital image must be extracted:

a. Image metadata.
b. Image thumbnail. It is the reduced version of the image to be analyzed, and it is at the header of the image file.
c. Camera traces. Footprints of the demosaicing, which is used to complete the pixels of the image when is created.
d. Compression signatures. Evidences in the image file of some image editing software.

iii) **Registration.** Image's technical features must be registered according with the chain of custody of the process. This registration must contain date, time and responsible of the extraction, as well as a brief description of the found technical features.

2.3 Step 3: Analysis of the image's technical features.

In this step the four technical features extracted from the digital image are analyzed in order to authenticate the digital image. For this purpose, the following premises must be considered as the analysis objects:

i) **Analysis of image metadata.** It is considered that when a digital image is modified, it may lose some metadata generated at the time of its capture. Thus, it is important to verify if the digital image preserves the metadata generated at the time of its capture. The image metadata considered are: brand and model of the camera, compression by software, orientation, date/time of capture and orientation of the image thumbnail.

ii) **Analysis of image thumbnail.** There are many software programs for image processing which are used to modify digital images, but these programs do not necessarily modify the image thumbnail. In this way, a thumbnail should be generated from the analyzed image, and then it must be compared pixel by pixel with thumbnail in the metadata file. Both must have the same dimensions. If the difference among the image thumbnails pixels is not significant (when at least the 90 percent of the thumbnails pixels are equals), it is defined that the image was not modified, but if the difference among them is significant, it is defined that the image was modified.

iii) **Analysis of camera traces.** This activity intends to verify the integrity of digital images and to detect the traces of tampering without using any protecting pre-extracted or pre-embedded information at the analyzed image [20]. When a digital image is captured, the camera makes an interpolation processing denominated demosaicing in order to complete the intensity values (pixels) of the digital image. This process affects the resolution and quality of the digital image. Thus, if the image has been modified, it is possible to find inconsistencies at the plane Y on the digital image, assuming that the color space is YCrCb. Plane Y suffers less loss of information when the JPEG compression is applied and the affectation by modification can be detected.

iv) **Searching compression signatures.** This activity intends to detect when an image editing software was used to make some change in a digital image. Regularly editing software leaves a compression signature in the header of that digital image.

At all time, the chain of custody must be considered in the step 3, and the registration of the hash values calculated when each action taken is performed.

Subsequently, it must be issued a technical resolution of each technical image's feature analyzed. In this resolution, it must be indicated whether the image approves or disapproves the testing. Finally a global technical resolution must be emitted to determine the image authenticity.

2.4 Step 4: Issuance of the dictum.

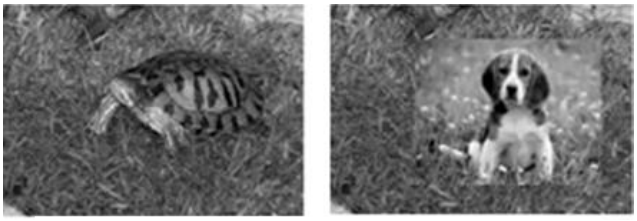
A dictum (verdict) that summarizes the conclusions of the analysis must be issued. This dictum must contain the image name, analysis date, image format, make and model of the camera used to capture the digital image, hash value of the image, brief description of the analysis performed, name of the analyst, results obtained at each step, and final technical resolution which indicates if the analyzed digital image is authentic, post produced or modified.

3. Technical application of the proposed methodology

In order to show the results of the application of this methodology, the following tools are going to be used: *Exiftool* to extract metadata, *Jhead* to extract the thumbnail image associated with the digital image, *JForensicsPG 1.0* an own software developed in Java 1.6 to generate a new thumbnail of the digital image and compare both thumbnails (extracted and generated); *JForensicsPG 1.0* is used for the camera traces analysis too, and *JPEGSnoop* that allows extracting the information of the header of the image, in order to verify the presence of any compression signature. Is important to mention that the software developed has intellectual property registration to the INDAUTOR, which is the organization that regulates the registration of software in Mexico. *JForensicsPG 1.0* has

the registration number 03-2012-022810521500-01 dated March 15, 2012.

For the example of the analysis of the image's technical features, two images are used. The first one image is named ORIGINAL.jpg, which is an image in the same state as it was generated at the time of its capture (with no modifications) with a device SAMSUNG GT-S5670L. The second one image is named MODIFICADA.jpg, which is an image modified with *Picasa 3* software; for generate MODIFICADA.jpg there was included in ORIGINAL.jpg a cut of another image. Fig. 1 shows the images used for the example.



a) *b)*

Fig. 1 Images used for the application analysis example; *a)* Image *ORIGINAL.jpg*, b) Image *MODIFICADA.jpg*.

3.1 Analysis of image metadata

In this analysis the following metadata are extracted: i) camera's make, ii) camera's model, iii) software compression, iv) image's orientation, v) date/time of image capture and vi) thumbnail's orientation. If it is possible to obtain at least four of these six metadata, this step is going to be approved; otherwise the result is going to be disapproved. Fig. 2 shows an example of the metadata extracted from an image without any change (ORIGINAL.jpg) using *ExifTool*.

```
ExifTool Version Number         : 7.38
File Name                       : ORIGINAL.jpg
Directory                       : C:\Users\Frank_07\Desktop
File Size                       : 1411 kB
File Modification Date/Time     : 2012:01:19 12:32:46
File Type                       : JPEG
MIME Type                       : image/jpeg
Exif Byte Order                 : Little-endian (Intel, II)
Make                            : SAMSUNG
Camera Model Name               : GT-S5670L
Orientation                     : Horizontal (normal)
Software                        : Imagen Digital ACD Systems
Modify Date                     : 2012:01:19 11:32:44
Y Cb Cr Positioning             : Centered
Exposure Time                   : 1/40
F Number                        : 2.6
Exposure Program                : Aperture-priority AE
ISO                             : 100
Exif Version                    : 0220
Date/Time Original              : 2011:10:29 18:10:53
Create Date                     : 2011:10:29 18:10:53
Max Aperture Value              : 2.6
Metering Mode                   : Center-weighted average
Flash                           : No Flash
Focal Length                    : 3.8 mm
User Comment                    : User comments
```

Fig. 2 Metadata of the image *ORIGINAL.jpg*.

Figure 3 shows the metadata obtained of MODIFICADA.jpg using *ExifTool*. The only change made was pasting a cut of another image with *Picasa 3* software. There is observed that make, model, orientation, software and date/time of capture are not found any more when the change was made in ORIGINAL.jpg. Therefore, for MODIFICADA.jpg result of this point is disapproved.

```
ExifTool Version Number         : 7.38
File Name                       : MODIFICADA.jpg
Directory                       : C:\Users\Frank_07\Desktop
File Size                       : 24 kB
File Modification Date/Time     : 2011:10:29 19:14:04
File Type                       : JPEG
MIME Type                       : image/jpeg
JFIF Version                    : 1.01
Resolution Unit                 : None
X Resolution                    : 1
Y Resolution                    : 1
Exif Byte Order                 : Little-endian (Intel, II)
Image Width                     : 240
Image Height                    : 320
Encoding Process                : Baseline DCT, Huffman coding
Bits Per Sample                 : 8
Color Components                : 3
Y Cb Cr Sub Sampling            : YCbCr4:2:0 (2 2)
Image Size                      : 240x320
-- press any key --
```

Fig. 3 Metadata of the image *MODIFICADA.jpg*.

3.2 Analysis of image thumbnails

For this action, it is necessary to extract the thumbnail associated to the image which is being analyzed and then a new thumbnail from this image must be generated. Both thumbnails must have the same dimensions. Then, the thumbnails (extracted and generated) must be compared pixel by pixel (considering the same position pixels comparison). For each pair of pixels compared, the difference should not exceed the absolute value of 8. This value was chosen as a maximum difference because it does not represent a significant change in the color perception of the human eye (considering this premise for 8-bit images). If more than ten percent (10%) of pixel differences vary for more than the absolute value of eight, the result of this phase is disapproved, otherwise it is approved. Fig. 4 shows graphically this comparison process:

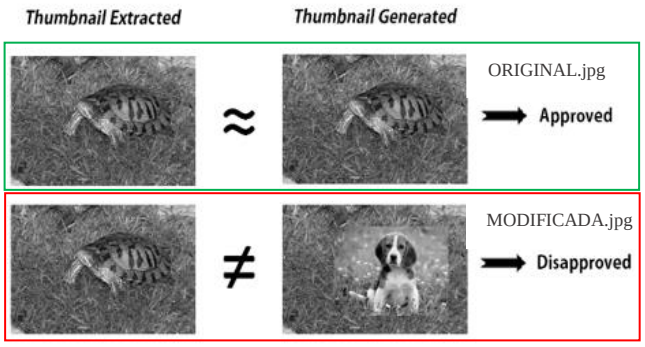


Fig. 4 Graphical thumbnails comparison example.

3.3 Analysis of the camera traces

This process analyzes 3×3 blocks of pixels of the Y plane of the digital image as proposed in [21]. This process calculates the values that must be generated in the demosaicing process for each 3×3 block, starting from left upper corner, up to the right bottom corner. The way to find the corner values of 3×3 blocks must be performed by increments of 2. For example, the position of the upper right corner of the first block is (0, 0), for the first horizontal offset, the position of the upper right corner of the second block would be (2, 0), the upper right corner of the third block would be (4, 0), and subsequently up to the end of horizontal blocks. Then, vertical position will be increased in 2 and horizontal position is reset (0, 2) to begin with horizontal offsets in the same way, up to up to go entirely through the Y plane in horizontal and vertical way. If the plane Y is not a multiple of 3, only must be gone up to the last position in which it is possible to extract an entire 3×3 block of pixels, i.e., it may not be scrolled maximum the last 2 lines of pixels, either horizontally or vertically.
Through this process, the four values at the corners of each block are extracted and then the remaining five values of the block are calculated. Notice in Fig. 5 that the black numbers correspond to the values of the corners of each block and numbers in gray are the values calculated.

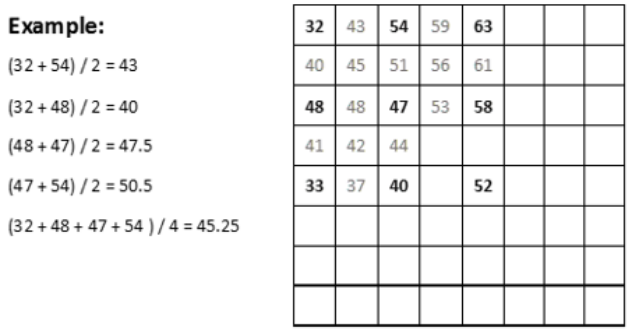


Fig. 5 Example of how a digital camera makes the interpolation process (demosaicing).

Then, the calculated values are compared with the values of the same position in the plane Y of the image. As in the thumbnail comparison, the difference among same position pixels of both planes does not have to be greater than the absolute value of 8. The reason for establishing this difference is because this is a non-significant difference in the color perception of the human eye. Subsequently, the following metric must be applied: If more than ten percent (10%) of pixel differences vary for more than the absolute value of eight, the result of this phase is disapproved, otherwise it is approved.

3.4 Searching Compression Signatures

In this action, the header of the image is reviewed in order to find a compression signature using an image editing software. If it is found a signature, the result of this searching is going to be disapproved; and in the final technical resolution the name of the software used for image processing must be printed. If it is not found a signature, the result is going to be approved. An example of the application of this point using the open source tool *JPEGSnoop* is shown in Fig. 6, where a signature of Adobe Photoshop software was found in the header of the image.

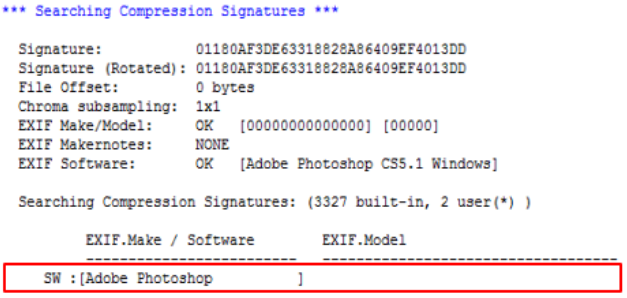


Fig. 6 Compression signature found using JPEGSnoop tool.

Therefore, the process applied to the image must be written in the chain of custody; it is important to obtain one more time the hash value (SHA 256) of the image used for the analysis. Then, this hash value must be compared with the hash value obtained before starting the analysis process. If the hash values compared are exactly equals, the process concludes successfully, otherwise, the process failed because of the management of the image during the process, and it cannot be used as evidence. Both cases (the one that applies) must be registered in custody chain.

4. Final dictum

The guide [13] indicates that in order to accept digital media as evidence, this media has to maintain the properties that authenticate it. Therefore, according to the process of analysis performed with the four technical features described above, the final technical resolution (verdict or dictum) that determines the image's authenticity is defined as follows:

i) If the fours kinds of analysis are approved, the final technical resolution is: ***Authentic***.
ii) If the first three kinds of analysis are approved and a compression signature of any image editing software is found, the final technical resolution is: ***Post Produced Image***. In this case, the image maintains the properties

that authenticate it, and the presence of that signature only indicates that the image has a quality improvement.

iii) Any other combination than the mentioned above, the final technical resolution is: ***Modified***.

Finally, a final dictum (verdict) must be emitted. This dictum must include the time and date of analysis, the name of the analyst, the name and the hash value (SHA 256) of the image analyzed, and a brief description of analyzed aspect including their respective result and the final technical resolution.

5. Results

For testing this methodology, a set of 450 digital images were used as follows: 150 original (without modifications), 150 modified (changing the fact that the image represents) and 150 post produced, only with quality improvement. The digital images used were generated with 10 different cameras of the following make and models of mobile devices: BlackBerry Curve 8530, Apple iPhone 4, Apple iPhone 5, Huawei Speed U8667, Nokia 3710 fold, Samsung GT-I8190, Samsung GT-I9300, Sony Xperia S LT26i, Sony Xperia U ST25i, Sony Ericsson Xperia Mini Pro HD SK17a. There were captured 45 images of each camera and divided as 15 originals, 15 modified and 15 post produced images. The smaller digital image has a size of 1600×1200 pixels which capture with a BlackBerry 8530 mobile device; and the bigger digital image has a size of 4000×2250 pixels captured with Sony Xperia S mobile device; the size of the images captured with the others 8 cameras are among that range. For modifying and post produce the images, the following software was used: *Picasa 3, Adobe Photoshop* and *Paint* of Windows XP.
Considering the sample of *N=450* digital images, divided in 150 originals images (O), 150 modified images (M) and 150 post produced images (P), three variables were defined, O_A, M_A and P_A in order to find the success verdict and determine the methodology efficiency. These variables indicate the times that the final technical resolution successful correspond to the group of the image analyzed (original, modified or post produced); in other words, these variables represent the number of image in the group minus the quantity of false negative in the technical resolutions obtained. These variables are defined according with Eqs. (1), (2) and (3).

$$O_A = O - F_{NO}\,, \tag{1}$$

$$M_A = M - F_{NM}\,, \tag{2}$$

$$P_A = P - F_{NP}\,, \tag{3}$$

where F_{Ni} is the amount of false negatives; *i* stands for *O* for Original, *M* for Modified or *P* for Post produced images.

In this way, the methodology efficiency is defined by Eq. (4):

$$E = \frac{1}{N}(O_A + M_A + P_A)\,, \tag{4}$$

where *N = 450* is the total number of digital image considered in the analysis.

The results obtained by applying the process described in this work are shown in Table 1, where the efficiency percentage was calculated, having a result of %E = 93.76.

Table 1: Results obtained by applying the proposed methodology

Image	**Detection**			**Total of detected digital images**	**Final resolution per 1/*N***
	O_A	**M_A**	**P_A**		
Original	142	1	12	155	0.3155
Modified	3	145	5	153	0.3222
Post Produced	5	4	133	142	0.2955
	150	**150**	**150**	**450**	***E*=0.9376**

Results in Table 1 show that there are also final technical resolutions with false positive. False positives are defined as the times that an authentic verdict was made when the Modified (*M*) or Post Produced (*P*) images were analyzed. Therefore, it is defined that false positives depend on false negatives of the other two groups of images. With this basis, Eqs. (5), (6) and (7) define the false positives where *i* stands for *O* for Original, *M* for Modified or *P* for Post produced images.

$$F_{PO} = F_{NM}k_3 - F_{NPP}k_1\,, \tag{5}$$

$$F_{PM} = F_{NO}k_2 - F_{NPP}(1-k_1)\,, \tag{6}$$

$$F_{PP} = F_{NO}(1-k_2) + F_{NM}(1-k_3)\,, \tag{7}$$

where F_{Ni} is the amount of false negatives and *i* stands for *O* for Original, *M* for Modified or *P* for Post Produced images.
These results also show that it is possible to calculate the values of false positives and false negatives of each group of images. Table 2 shows these particular calculations:

Table 2: False positives and false negatives found in the results shown in Table 1

Groups of images	False Negatives	False positives
Original	8	13
Modified	5	8
P. Produced	17	9

In Eqs. (5), (6) and (7), it can be observed that there are three factors k_1 , k_2 and k_3 . These factors indicate a specific weight of the false negative results of one group of images, which directly affect the quantity of false positive results of the other groups. Clearing these three factors in Eqs. (5), (6) and (7), are obtained Eqs. (8), (9) and (10).

$$k_3 = \frac{F_{PO} - F_{NPP}k_1}{F_{NM}}, \qquad (8)$$

$$k_1 = 1 + \frac{(-F_{PM} + F_{NO}k_2)}{F_{NPP}}, \qquad (9)$$

$$k_2 = 1 + \frac{(-F_{PP} + F_{NM} - F_{NM}k_3)}{F_{NO}}, \qquad (10)$$

Finally, solving Eqs. 8, 9 and 10 by substituting the results of the false positives and false negatives of Table 2, the values of the weight factors are obtained:

$$k_1 = \frac{9}{17},\ k_2 = 0 \quad \text{and} \quad k_3 = \frac{4}{5}.$$

These factors are useful when it is necessary to calculate the false positive values of other groups of images, where the distribution of the images is unknown.

6. Conclusions

It was feasible to define a methodology which determines if a digital image in JPEG format is authentic, post produced or modified, based on internationally accepted guides and best practices about evidence management. The process applied to metadata, thumbnail, camera traces and compression signatures found in the digital image provides a robust analysis that grants certainty and reliability of the dictum emitted. The efficiency percentage obtained is 93.76% of the proposed methodology; therefore, it can be applied to help in the clarification of facts or events arising from security incidents with legal, civil, administrative or criminal implications. According to laws of each country, this process may help to present a digital image as evidence.

The proposed methodology can be applied using open software tools, like is shown in the technical application example. However, it is possible to develop software that automatizes the four analysis process (metadata, thumbnail, camera traces and compression signatures), because each aspect of analysis is in a digital way and only requires computer processing.

Acknowledgments

R. Vázquez-Medina wishes to thank Instituto Politécnico Nacional (IPN México) for financially support this research through grant SIP/IPN 20150316 and SIP-2015-RE/013. F. Rodríguez-Santos (CVU-377075), G. Delgado-Gutiérrez (CVU-372164) and L. Palacios-Luengas (CVU-373990) thank for the scholarship provided by CONACYT–México. Technical and computational support from J. L. Pichardo- Méndez (CVU-668444) is gratefully acknowledged.

References

[1] GRABLER F, AGRAWALA M, LI W, DONTCHEVA M, IGARASHI T. Generating photo manipulation tutorials by demonstration. ACM Transactions on Graphics (SIGGRAPH) 28, 3, 2009.

[2] GARRY M, GERRIE M. When photographs create false memories. Current Directions in Psychological Science 14, 326–330, 2005.

[3] JOHNSON M, FARID H. Detecting photographic composites of people. In 6th International Workshop on Digital Watermarking. Guangzhou, China, 2007.

[4] FARID H. A survey of image forgery detection. IEEE Signal Processing Magazine 2, 26, 16–25, 2009.

[5] FARID H, BRAVO M. Image forensic analyses that elude the human visual system. In SPIE Symposium on Electronic Imaging. San Jose, CA, 2010.

[6] KEE E, FARID H. Exposing digital forgeries from 3-D lighting environments. In Workshop on Information Forensics and Security, 2010.

[7] O´BRIEN JF, FARID H. Exposing Photo Manipulation with inconsistent reflections. ACM Transactions on Graphics. 31(1):4:1–11, 2012.

[8] LUO WEIQI, QU ZHENHUA, PAN FENG, HUANG JIWU. A survey of passive technology for digital image forensics. Front. Comput. Sci. China, 1(2): 166−179, 2007.

[9] MIN-GU HWANG AND DONG-HWAN HAR. A Novel Forged Image Detection Method Using the Characteristics of Interpolation. J Forensic Sci., Vol. 58, No. 1, January 2013.

[10] FEI PENG, JIAO-TING LI, MIN LONG. Identification of Natural Images and Computer-Generated Graphics Based on Statistical and Textural Features. J Forensic Sci., Vol. 60, No. 2, March 2014.

[11] MIN GU HWANG, DONG HWAN HAR. Identification Method for Digital Image Forgery and Filtering Region through Interpolation. J Forensic Sci., Vol. 59, No. 5, September 2014.
[12] YANJUN CAO, TIEGANG GAO, GUORUI SHENG, LI FAN, LIN GAO. A New Anti-forensic Scheme — Hiding the Single JPEG Compression Trace for Digital Image. J Forensic Sci, Vol. 60, No. 1, January 2015.
[13] KENT K, CHEVALIER S, GRANCE T, DANG H. Guide to Integrating Forensic Techniques into Incident Response. National Institute of Standards and Technology. Special Publication 800-86. August 2006.
[14] Association of Chief Police Officers. Good Practice Guide for Computer-Based Electronic Evidence. Official release version.
[15] A Road Map for Digital Forensic Research. Report from the First Digital Forensic Research Workshop (DFRWS) Technical Report; 2001 Nov. DTR -T001-01 Final.
[16] CICHONSKI P, MILLAR J, GRANCE T, SCARFONE K. Computer Security Incident Handling Guide. National Institute of Standards and Technology. Special Publication 800-61 Revision 2. August 2012.
[17] Scientific Working Group on Digital Evidence. Best Practices for Computer Forensics v.3.1. September 2014.
[18] National Institute of Justice. Forensic Examination of Digital Evidence: A Guide for Law Enforcement. NCJ 199408.
[19] National Institute of Justice. Electronic Crime Scene Investigation: A Guide for First Responders, Second Edition. NCJ 219941.
[20] BABAK MAHDIAN, STANISLAV SAIC. A bibliography on blind methods for identifying image forgery. Signal Processing: Image Communication, Volume 25, Issue 6, July 2010.
[21] Farid H. 5 ways to spot a fake photo. The Scientific American Magazine; Page 5, June 2, 2008.

Francisco Rodríguez-Santos Obtain the Engineering degree in 2010 and subsequently the master in information security degree in 2012. Nowadays, studies the PhD. in the Section of postgraduated studies and research which belongs to the Superior School of mechanic and electric engineering of Instituto Politécnico Nacional in Mexico City. He is a member of the Forensics workgroup of Mexican Accreditation Entity and his interest areas are the information security, information forensics, pattern recognition and regulatory compliance in information security.

Guillermo Delgado-Gutierrez Master in Microelectronics Engineering graduated in 2013 in the Section of postgraduated studies and research which belongs to the Superior School of mechanic and electric engineering of Instituto Politécnico Nacional in Mexico City. Nowadays, studies the PhD. in the same Section of Studies. He is a member of the Forensics workgroup of Mexican Accreditation Entity and his interest areas are information security, signal processing, digital forensics analysis and regulations.

Leonardo Palacios-Luengas Master in Microelectronics Engineering graduated in 2012 in the Section of postgraduated studies and Rresearch which belongs to the Superior School of mechanic and electric engineering of Instituto Politécnico Nacional in Mexico City. Nowadays, studies the PhD. in the same Section of Studies. His interest areas are cryptography, steganography, embedded systems programming and digital electronics design.

Rubén Vázquez-Medina He obtained the Electronics Engineering degree in 1988, he has a Master in Sciences of Electric Engineering, obtained in 1993. He has the PhD. Degree obtained in 2008 in the Universidad Autónoma Metropolitana. Since 2006 is a guest professor in the Master in Information Security program of the Superior Naval Studies Center of Mexican Navy. Nowadays is the subdirector of the postgraduated section of the Cleaner Production Mexican Center since 2014. He conducts research in the areas of cryptography, steganography, computer forensics and compliance in information security; and modeling diffusion reaction systems for applications in information security.