Natarajan Meghanathan
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

The Sixth International Conference on Wireless & Mobile Networks
(WiMoNe - 2014)
Sydney, Australia, December 27 ~ 28 - 2014

AIRCC

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-18-2
DOI : 10.5121/csit.2014.41201 - 10.5121/csit.2014.41223

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Sixth International Conference on Wireless & Mobile Networks (WiMONe 2014) was held in Sydney, Australia, during December 27 ~ 28, 2014. Sixth International Conference on Network & Communications Security (NCS 2014), International Conference on Signal, Image Processing and Multimedia (SPM-2014) and The International Conference on Computer Science, Engineering and Information Technology (CSEIT-2014). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The WiMONe-2014, NCS-2014, SPM-2014, CSEIT-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, WiMONe-2014, NCS-2014, SPM-2014, CSEIT-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the WiMONe-2014, NCS-2014, SPM-2014, CSEIT-2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai

Organization

General Chair

David C. Wyld	Southeastern Louisiana University, USA
Jan Zizka	Mendel University in Brno, Czech Republic

Program Committee Members

Ahmed Nabih Zaki Rashed	Menoufia university, Egypt
Ahmed Y. Nada	Al-Quds University, Palestine
Ali Azimi	Ferdowsi university of Mashhad, Iran
Ali Poorebrahimi	Islamic Azad University, Iran.
Alireza Afshari	Islamic Azad University, Iran
Amani Samha	Queensland University of Technology, Australia
Amir Hosein Jafari	Iran University of Science and Technology, Iran
Ankit Chaudhary	Maharishi University of Management, USA
Assem Abdel Hamed Mousa	E commerce tech support systems, Egypt
Ayad Ghany Ismaeel	Erbil Polytechnic University, Iraq
Ching-Nung Yang	National Dong Hwa University,Taiwan
Cristina Alcaraz	University of Malaga, Spain
Dac-Nhuong Le	Vietnam National University, Vietnam
Diallo Abdoulaye	ServiceRocket, Malaysia
Diganta Saha	Jadavpur University, India
Dires, Fasil Fenta	University of Gondar, Ethiopia
Edward Ofoegbu	Oduduwa University, Nigeria
Farideh Alizadeh	University of Malaya (UM), Malaysia
Fernando Bobillo	University of Zaragoza, Spain
Francine Krief	University of Bordeaux, France
Habil Bencsik Andrea	Szechenyi Istvan University, Hungary
Haider N. Hussain	Handover in wireless network, Iraq
Hamid Mcheick	Universite De Sherbrooke, Canada
Hassini Noureddine	University of Oran, Algeria
Hicham Behja	University Hassan II Casablanca, Morocco
Himanshu Aggarwal	Punjabi University, India
Ikram Uddin	University of Haripur, Pakistan
Isa Maleki	Islamic Azad University, Iran,
Islam Atef Abd-elgawad	Alexandria University,Egypt
Islam Atef	Alexandria University, Egypt
Israa SH.Tawfic	Gaziantep university,Turkey
Justinian Anatory	University of Dodoma, Tanzania
Keneilwe Zuva	University of Botswana, Botswana
Khaled Saad Eldin Mohamed Ragab	National Research Center, Egypt
Kannan A	Anna University, India
M.Naderinejad	Tehran University of Medical Sciences, Iran
Mahdi Mazinani	Azad University, Iran

Manoj Vasanth Ram	Advanced Micro Devices, USA
Mohammed Erritali	Sultan Moulay Slimane University, Morocco
Nabila Labraoui	University of Tlemcen, Algeria
Naderinejad M	Tehran University of Medical Sciences, Iran
Natarajan Meghanathan	Jackson State University,USA
Peiman Mohammadi	Islamic Azad University, Iran
Rahil Hosseini	Azad University,Iran
Rajkumar Patro	Haramaya University, Ethiopia
Reda Mohamed HAMOU	Tahar Moulay University of Saïda, Algeria
Saad M. Darwish	Alexandria University,Egypt
Saeed M Agbariah	George Mason University, USA
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Seyyed Reza Khaze	Islamic Azad University, Iran
Shandilya V	University of Memphis, USA
Subarna Shakya	Tribhuvan University, Nepal
Vasanth Ram Rajarathinam	Advanced Micro Devices, USA
Vimal Kumar	University of West Florida, USA

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Sixth International Conference on Wireless & Mobile Networks (WiMoNe - 2014)

Sixth International Conference on Network & Communications Security (NCS - 2014)

International Conference on Signal, Image Processing and Multimedia (SPM-2014)

The International Conference on Computer Science, Engineering and Information Technology (CSEIT-2014)

ENERGY-BALANCED IMPROVED LEACH ROUTING PROTOCOL FOR WIRELESS SENSOR NETWORKS

KHAMISS.A.A, CHAI Senchun, ZHANG Baihai and LI Qiao

School of Automation, Beijing Institute of Technology, Beijing, China
Ahmed.abdo7728@yahoo.com, Chaisc97@bit.edu.cn,
smczhang@bit.edu.cn,442302839@qq.com

ABSTRACT

A proper sensor node clustering is an effective topology control that can balance energy consumption among sensor nodes and increase network scalability and life time. As the use of wireless sensor networks (WSNs) has grown enormously, the need for energy-efficient routing and data aggregation has also risen. LEACH （Low Energy Adaptive Cluster Hierarchy）is a hierarchical clustering protocol that provides an elegant solution for such protocols. Random clustering is the main deficiency of LEACH. In this paper an energy balanced clustering approach is proposed, in which the K-mean clustering algorithm is applied. It is centralized clustering algorithm that based on minimum energy clustering to form optimal clusters. For the candidate nodes, the location and the residual energy are used as key parameters to select the cluster head (CH). The method shows that the proposed approach outperforms LEACH in terms of energy conservation and network life time prolonging.

KEYWORDS

K-mean, Role Factor, Cluster Energy & Root Node.

1. INTRODUCTION

WSNs typically consist of a large number of low-cost, low-power and multifunctional wireless sensor nodes, with sensing, limited communication and computation capabilities. These nodes communicate via a wireless medium and collaborate to accomplish a common task, such as environment monitoring, military surveillance and industrial process control [1]. The basic philosophy behind WSNs is that, while the capability of each individual node is limited, the aggregate power of the entire network is sufficient for the required mission [2]. In most applications, nodes are deployed randomly, and once deployed they must be able to autonomously organize themselves into a network. Generally WSNs are characterized with high density of deployed nodes, while the nodes themselves are unreliable and restricted in power, computation and storage resources. And due to these characteristics, suitable network routing protocols are required to implement various network control and management functions. A WSN typically has little or no infrastructure; and a great number of nodes allow sensing larger geographical region with greater accuracy. The collected data is sent usually via radio transmitter to the sink either directly or through a gate way. Routing protocol is one of the core technologies in WSNs that is full of challenge due to its inherent characteristics [3, 4]. Routing is the process of determining the best data transmission path between source and destination. Many protocols

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 01–11, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41201

have been proposed to forward the correct data to the sink and enhance the life time of the WSN [5]. Using clustering techniques in routing protocols prolongs the life time of the network and contributes in over all system scalability. Clustering is a well-known and widely used data analysis technique that is useful in applications which require scalability. Data transmission is the most critical energy consumption issue for sensor node, so, by clustering, communication can be single-hop or multi-hop. A single-hop communication can reduce the over head cost, but when communication distance increases, it consumes more energy. Multi-hop communication consumes less energy than single-hop for long distances, but it has more over head cost and delay. Then, multi-hop routing is suitable for applications with large scale networks. The routing protocols in WSNs can be classified into two categories according to application requirements, namely data-query based routing and data-gathering based routing [6]. The Energy-efficient routing forwards packets along the path with minimum energy, this leads to minimum energy consumption. But it causes unbalanced distribution of residual energy and nodes closer to sink deplete their energy faster than others. Such imbalance definitely shortens the life time of the network. And the connectivity between nodes and sink can be maintained for longer time if nodes consume their energy evenly. Therefore, balanced energy consumption must be considered beside energy-awareness when designing energy-efficient routing protocol. The energy balanced routing distributes the levels of residue energy evenly throughout the network, extending the network life time. Recently, many research activities have been done in designing energy-efficient protocols for WSNs. Most of them focus on finding minimum energy path for data transmission and neglect the survivability of the entire network. Minimum energy consumption schemes minimize energy consumption by forwarding data through few popular paths. Therefore, energy consumption on these paths increases quickly causing imbalance of residual energy throughout the network, decreasing the overall performance and life time of the network [7].

Generally, the main reasons that cause imbalance of energy distribution in WSNs can be summarized in:

- Topology: The initial deployment limits the number of paths through which data can flow.
- Application: The applications may determine the location and the rate at which the nodes generate data. So, the area generating more data and the path forwarding more data may deplete energy faster.
- Routing: Minimizing energy consumption via using static optimal path results in energy imbalance because energy of optimal path nodes is quickly depleted.

According to the above reasons, some possible solutions had been suggested to balance energy consumption:

- Deployment optimization: This will solve the problem of mismatch between topology and application according to the traffic pattern.
- Topology control: Nodes collaboratively adjust their transmission power to form a proper network topology.
- Relay nodes: Relay nodes can relieve heavily loaded areas or paths in a way similar to optimization of deployment.
- Data aggregation: It exploits redundancy to minimize energy consumed in data transmission.
- Energy-Balanced Routing: It maintains the network connectivity for a longer time and prolongs the entire network life time [8].

WSNs are resource constrained, so, such networks need novel protocols and algorithms which can utilize the available resources optimally and meet user requirements. The above points, in

addition to energy-efficiency and balanced energy consumption are important issues when designing energy-efficient routing protocols for WSNs. In this paper we proposed a routing scheme that overcomes the problem of energy imbalance in LEACH routing protocol. It demonstrates the advantages of balanced energy consumption across the network. The remainder of this paper is organized as follows: The next section describes the related works and section III explains the system model. Section IV demonstrates the proposed approach while section V contains the results obtained; and finally section VI represents the conclusion of the work.

2. RELATED WORKS

Routing in WSNs is very challenging due to the inherent characteristics that distinguish the WSNs from other wireless networks like ad hoc or cellular networks. The aim of network layer is to route data from sensor to sink in an energy-efficient and reliable manner in order to extend the network life time. Many research activities had addressed the energy-efficient routings; most of them focused on minimizing energy consumption on local nodes or the entire network via finding an optimal path. In [2], some modified versions of LEACH had been discussed. LEACH is a cluster-based protocol that randomly selects CHs, where the number of CHs is adaptive and the selection process is dynamically rotated to evenly distribute load [9]. LEACH suffers from random selection of CHs and it works best only if the energy of nodes is uniform. There is no any certainty about the distribution of CHs throughout the network, and the idea of dynamic clustering brings extra overhead. Moreover, it is not applicable for large scale applications. E-LEACH, improves the CH selection by making the residual energy of the node as the main metric which decides whether the nodes turn into CHs or not after the first round. In the first round all nodes have the same probability to become CH, i.e. randomly as LEACH, in the next rounds the residual energy of each node is different and it is taken into account. That means, nodes have more energy will become CHs rather than nodes with less energy [10]. TL-LEACH (Two-level Hierarchy LEACH) is enhanced version of LEACH protocol; it has two levels of CHs. In this protocol, CH collects data from other cluster members as original LEACH, but rather than transfer data to the BS directly, it uses one of the CHs that lies between the CH and the BS as a relay station [11].TL-LEACH faces the same problem as LEACH, since it uses the same mechanism. LEACH-C is a centralized clustering algorithm; it produces better clusters by dispersing the CHs evenly through the network. During the setup phase of LEACH-C each node sends information about its current location and residual energy to the sink. Sink node computes the node average energy and determines which nodes have energy below this average [9]. In LEACH, each CH directly communicates with the BS no matter the distance between CH and BS. On the other hand M-LEACH (multi-hop LEACH) selects the optimal path between the CH and BS through other CHs and uses the CHs as a relay station [12]. M-LEACH is almost the same as LEACH, only makes communication mode between CHs and BS from single hop to multi hop. In LEACH-H (Hybrid cluster head selection LEACH) the base station selects the CHs in first round. While in the followed rounds the CHs select the new CHs in their own clusters. Hybrid Energy-Efficient Distributed Clustering (HEED) is a multi hop clustering that focus on efficient clustering based on the physical distance between nodes. The scheme selects CHs in terms of residual energy and intra-cluster communication cost, which is useful if given node falls within the range of more than CH [13]. The problem in HEED is that, it is important to identify what is the range of a node in terms of its power level, as node will has multiple discrete transmission power levels. In [14], FZ-LEACH (Far-Zone LEACH) is proposed to form far-zone which is a group of nodes which are placed at locations where their energies are less than a threshold. In ACHTH- LEACH (Adaptive Cluster Head Election and Two-Hop LEACH) protocol [15], nodes are tagged as near or far nodes according to the distances to the BS. The near nodes belong to one cluster, while the far nodes are divided in to different clusters. The node with the maximal residual energy in each cluster is selected as CH.

In this paper we propose scheme that provides how to prolong the network life time and make an efficient use of the critical resources located at the nodes. We aim to create more intelligent clusters, minimize the number of nodes in a cluster, and select only a single CH in each cluster and not to affect the communication functionalities. The scheme adopted minimum energy clustering technique to form clusters and minimum communication cost metric to select CH in each cluster. Also minimum energy path multi-hop routing had been used to transmit data to the base station.

3. SENSOR RADIO MODEL

3.1. Assumptions

The following assumptions are made about the sensors and the network model:
1-The sink node is located inside the sensing field.
2-Sensors are location-aware; energy constrained
3-Sensors are energy-aware; energy resource-adaptive.
4-Links are symmetric; nodes and sink both are stationary
5-Cluster: a mixture of single and multi-hop, no assumption about homogeneity of node category and homogeneity of dispersion radius.

3.2. Sensor Energy Model

We adopt he energy consumption and radio model given in [9]. To achieve an acceptable signal-to-noise ratio (SNR) in transmitting k bit message over a distance d , the energy cost of transmission (E_{TX}) is given by:

$$\begin{cases} E_{TX}(k,d)=kE_{elec}+k\varepsilon_{fs}d^2; \text{ if } d<=d_0 \\ E_{TX}(k,d)=kE_{elec}+k\varepsilon_{mp}d^4; \text{ if } d>d_0 \end{cases} \quad (1)$$

Where E_{elec} = 50nj is the energy utilized to run the transceiver circuit, $\varepsilon_{fs}=10j/bit/m^2$ and $\varepsilon_{mp}=0.0013pj/bit/m^4$ are energies utilized by transmission amplifiers for short and long distances respectively. For $d=d_0$, the distance threshold is given by, $d_0=\sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$ where d_0 is known as the crossover distance and $E_d=5nj/bit/signal$ is the fusion energy per bit. To receive a k bit message, the radio expends receiving energy (E_{RX}) which is given by:

$$E_{RX}=kE_{elec} \quad (2)$$

4. PROPOSED ALGORITHM

The proposed scheme is a cluster-based algorithm that modifies the clustering process. Thus, both the cluster formation and the CH selection methods had been modified. For more energy conservation, a multi-hop routing was adopted to forward data to the BS. Clustering achieves an important improvement in terms of energy consumption, and it is crucial in scaling the networks.

Also clustering is one of the basic approaches to design energy-efficient distributed WSNs. However, these benefits can result in extra overhead due to cluster formation's message exchange. In this scheme we assumed exchanging sensor data may be an expensive network operation, but exchanging data about sensor data needs not be. Clustering can be performed in centralized way by BS (as in our scheme) or in distributed way where every node decides autonomously about its role. Furthermore, cluster formation can be either static or dynamic according to whether the network is heterogeneous or homogenous. Since energy is the major concern, then balanced energy consumption is important in energy conservation. The operation of the proposed scheme is divided into rounds; each round consists of setup phase and data transmission phase. The cluster formation, CHs selection, and multi-hop paths establishment are done successively in setup phase. While in data transmission phase, the nodes sense and transmit data to CHs and then CHs forward data to BS after aggregation. In LEACH, the cluster heads are selected according to formula (3) and then members joint their nearest CHs to construct clusters in setup phase. In the proposed scheme, clusters are formed firstly, and then a single cluster head is selected in each cluster.

$$T(n)=\begin{cases}\dfrac{p}{1-p*(r \bmod(\frac{1}{p}))} & if n\in G \\ 0 & if n\notin G\end{cases} \tag{3}$$

Where p represents the percentage of CHs in the network, r is the current round and G represents the group of nodes those had not been selected as cluster head.

4.1. Setup Phase

4.1.1. Cluster Formation

Firstly, since the BS resources are not limited, the BS uses the unsupervised clustering technique K-mean Clustering method to divide the nodes into k clusters (k= $p \times n$). It can be viewed as a greedy algorithm for partitioning the total n nodes into k clusters so as to minimize the sum of squared distances to the cluster centers. The division based on network topology, nodes distribution and the similarity of criterion for grouping. For K-mean clustering method, the similarity criterion in our case is the geometrical distance. Each node is randomly assigned to a cluster, and then the cluster centers are computed. Repeating of this process produces clusters with minimum energy. Cluster energy is defined as the sum of the distance squared from each node to its cluster center. It is the goal of the K- mean algorithm to find, for fixed number of clusters, a clustering that minimizes this energy. The algorithm aims at minimizing an objective function, which is the squared error function;

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j-c_j\right\|^2 \tag{4}$$

Where $\left\|x_i^j-c_j\right\|^2$ represents chosen distance measure between a node x_i^j （node i in cluster j ） and cluster center c_j (center of cluster j), J is an indicator of the distance of the total n nodes from their respective cluster centers. The algorithm works as follows; in the beginning, determine the number of clusters k, the centers of these clusters (initial) might be assumed

randomly or the first k nodes can serve as the initial centers. K-mean algorithm will do the following three steps until converge (iterate until stable, i.e. no object move group):

1- Determine the centroids coordinate.
2- Determine the distance of each node to the centroids.
3- Group the nodes based on minimum distance.

Fig.1 summarizes the steps of k-mean. After partitioning, the BS notifies the nodes by sending a message containing the cluster ID, distance to cluster centroid, distance to BS and TDMA schedules for each cluster to be employed in steady state phase. Every node stores this message and uses its contents to decide whether it will turn into cluster head or not.

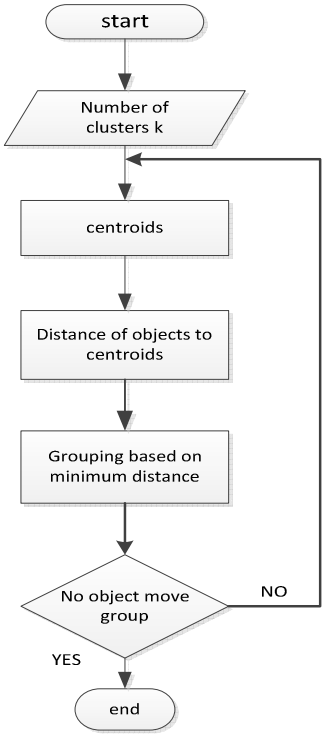


Fig 1. K-mean algorithm steps

4.1.2. Cluster Head Selection

Cluster Head Selection: In LEACH, CH selection based on the remaining energy, percentage of CHs in the network and the number of times the node has been CH. In the proposed algorithm, the CH selection equation of LEACH had been extended by inserting the residual energy and the distances to BS and cluster centroid. Thus, a node with more residual energy and close to both cluster center and BS will be more eligible to become a CH in each cluster. CH selection equation in (3) can be rewritten as:

$$T(n) = \frac{p}{1 - p * (r \bmod (\frac{1}{p}))} * \frac{E_{resi}}{E_{initial}} * \frac{1}{({d_{BS}}^2 + {d_{Cent}}^2)} \qquad (5)$$

Where E_{resi} represents the residual energy of node n, $E_{initial}$ stands for initial energy of all nodes (assuming that all nodes have equal initial energy), and d_{BS} and d_{Cent} are the distances of node n to the BS and cluster center respectively. In the proposed algorithm, the probability of being CH increases when a node has more residual energy and more closer to the BS and cluster centroid. Such situation gives a good chance to random number r chosen by the concerned node n to be less than the threshold value $T(n)$. Hence, in each group of nodes, those have the same centroid; the node that has the minimum value of r becomes a CH in that group. The CHs advertise themselves and their centroids, and then every member node belongs to its CH according to the cluster centroid and receives its TDMA schedule. The selected CH will continue to work as cluster head until it dies or another CH dies.

4.1.3. Establishing Multi Hop Paths

The proposed algorithm takes a mixture of single and multi-hop routing. At first, the BS computes and broadcasts the average distance d_{avg} to all nodes according to $d_{avg} = \frac{\sum_{i=1}^{N} d_{BS}}{N}$. If the distance between CH and BS is less than or equal to the average distance the CH communicates with BS directly. Otherwise, CH sends a message to other neighbor CHs including cluster head ID, distance to BS and residual energy. Neighbor CHs save this message at first, then feedback their own messages. Based on the feedback, the CH chooses a neighbor CH which has more residual energy and close to BS as its next hop node. Finally, all CHs find their parent nodes, and the parent node that directly communicates with BS is called root node. The root node sends data to BS directly without aggregation.

4.2. Data Transmission Phase

This phase consists of frames; in each frame all nodes send data to their CHs or BS. The CHs open their radios to receive data from nodes. Normal nodes are placed in to sleep mode, and every one opens its radio in its own time-slot. Every node saves its residual energy when it is sending data to CH or to the BS. When any cluster head dies, the BS informs the network nodes about the starting of the cluster reformation step. Then the data transmission phase is terminated and the entire system moves to the setup phase. Because when some nodes die, the total number of nodes decreases and the distribution of the nodes changes, hence, the number of cluster heads must be adjusted to keep the adaptivity of the algorithm. The system continues these rounds until every node's energy has been depleted.

4.3. Energy Analysis

Assuming $_N$ is the total number of nodes and K is the number of clusters; the average number of nodes per cluster is $\frac{N}{K}$. Each CH dissipates energy of receiving data from members, aggregating and transmitting the aggregated data to the BS. So, the energy dissipated by the CH during single frame is:

$$E_{CH} = lE_{elec}(\tfrac{N}{K} - 1) + lE_{DA}(\tfrac{N}{K}) + E_{elec}(c.\tfrac{N}{K}.l) + \varepsilon_{mp}(c.\tfrac{N}{K}.l)d_{Bs}^4 \qquad (6)$$

E_{CH} represents the total energy dissipated by CH during each frame, $lE_{elec}(\frac{N}{K}-1)$ is the energy dissipated for receiving data from members, $lE_{DA}(\frac{N}{K})$ is the aggregation energy and the transmitting energy to the BS is $E_{elec}(C.\frac{N}{K}.l)+\varepsilon_{mp}(C.\frac{N}{K}.l)*d_{Bs}^4$; C is the aggregation coefficient and l is the message length. On the other hand, member nodes need energy to transmit data to CH once during each frame, which can be estimated as:

$$E_{mem} = lE_{elec} + lE_{fs} * d_{CH}^2 \quad (7)$$

Where d_{CH}^2 is the distance between CH and member node in the same cluster. Therefore, the energy dissipated in a cluster during each frame is

$$E_{cluster} = E_{CH} + (\frac{N}{K} - 1) E_{mem} \quad (8)$$

And the total energy for one frame is:

$$E_{total} = KE_{cluster} \quad (9)$$

Then according to the number of frames in each round, the life time of the network can be calculated in terms of number of rounds. Number of rounds can be calculated by dividing the total network energy by energy expended in one round.

5. SIMULATION AND RESULTS

For simulation environment, 100 nodes are deployed randomly over an area of 100×100 meters, and the BS is located in the sensing field. The control and data message lengths are 200bits and 6400 bits respectively, with aggregation factor c set to unity as in (6), such that every node has data to send in each round. In this work the following metrics are used to evaluate the life time and performance of the network; time taken till all nodes die, data sent to BS and stability period which is the time till the first node die. Fig. 2 shows the live nodes versus simulation time, from the figure, in the proposed approach the number of rounds during simulation time is nearly twice the number of rounds in normal LEACH and the stability period is increased by 29%. This demonstrates the dramatic improvement in the network life time and performance.

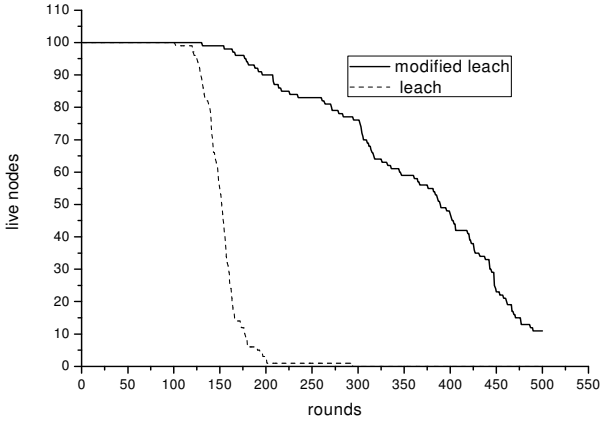


Fig 2. Live nodes over simulation time

This is due to the minimum energy clustering technique and the selection of the most suitable nodes as CHs. The two techniques minimize the cost in terms of energy while nodes are communicating with CHs or the later are communicating with BS. In fact the mission of each node depends on a factor that consists of the node energy, the minimum energy needed for that mission and the location of the node. This permits nodes to dissipate energy evenly, which lets nodes to stay alive for long time. Fig. 3 shows the data sent to BS during the simulation, the data sent to BS during the simulation time is increased by approximately 36% than normal LEACH. This is due to the fact that the network life time is increased in compare to LEACH, which demonstrates the significant improvement in the network performance with the proposed approach.

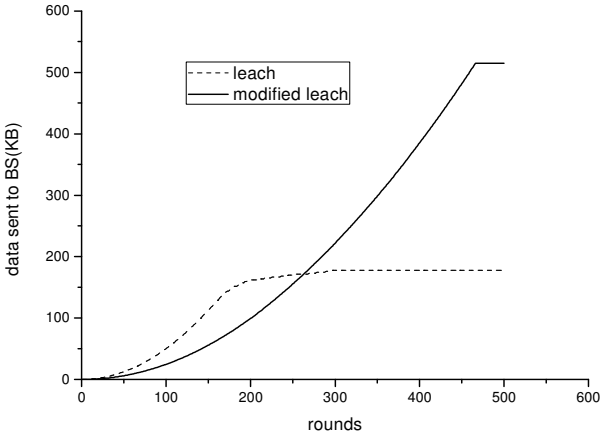


Fig 3. Data sent to BS over simulation time

Fig. 4 and Fig. 5 show the energy variation versus data sent and time respectively. The mission of each node depends on the role factor, so the heavy load mission will be done by nodes that have high role factor, while nodes with low role factor do sensing mission. This explains why the system stays stable for long time. Thus, in proposed approach the network does not exhaust energy suddenly and quickly. From the above results the minimum cluster energy method prolongs the stability period and life time of the wireless network, there by improves the efficiency of the network.

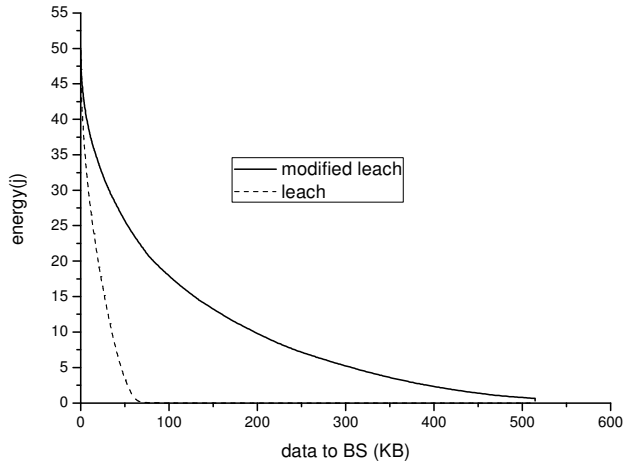


Fig 4. Energy over data sent to BS

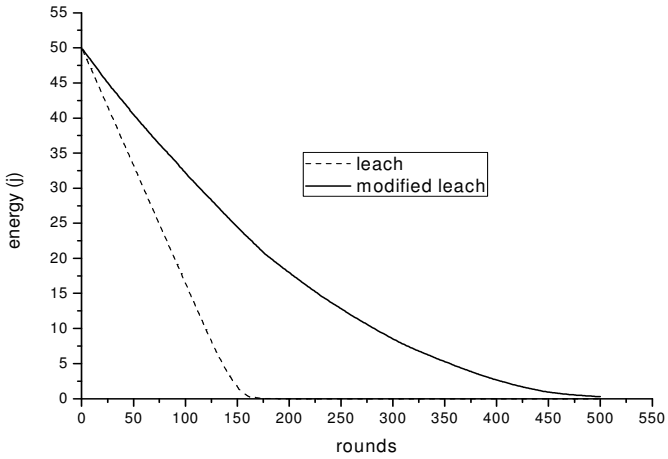


Fig 5. Energy over rounds

6. CONCLUSIONS

The transmission distance has a great impact on the energy consumption as well as the regular consumption of the energy over the entire network. Most energy-efficient protocols focus on minimum energy path, although balanced energy consumption has direct effect on energy efficiency. So, a technique that minimizes the cluster energy was used to form minimum energy clusters, in addition to energy-aware method to select the CH. Nevertheless, a minimum energy path is used to forward data in multi-hop routing to minimize energy consumption. The simulation results show that the algorithm can balance the load between nodes and prolong the stability period and life time of the network in comparison to LEACH. The approach does not require cluster formation in every round, but it depends on current CH state, this decreases the dynamic clustering overhead. Although the results confirmed that the proposed algorithm outperforms LEACH in lengthening WSN life time, there are many protocols that have to be

compared. Furthermore, the number of member nodes in clusters, the duration of the data transmission phase and whether k-mean will always terminate quickly and its complexity will be studied in future. And more factors that can affect the life time of WSN will be considered. It can be proved that K-mean will always terminate, but the algorithm does not necessary find the most optimal configuration, corresponding to the global objective function minimum. Also it is significantly sensitive to the initial randomly selected cluster centers, since it is iterative clustering algorithm. So, a "good" selection of initial cluster centers is an essential clustering problem.

REFERENCES

[1] K. Akkaya and M.Younis, (2005), "Asurvey of Routing Protocols in Wireless Sensor Networks", Elsevier Ad Hoc Network Journal, Vol. 3/3, pp. 325-349.

[2] C. Shilpa, Dr. M. Sona and C. Yogesh, (July 2012)"Asurvey of Hierarchical Routing Protocols in Wireless Sensor Networks", International Journal of Emerging trends in Engineering and Development, ISSN 2249-6194, Issue 2, Vol. 5

[3] Jennifer Yick, Biswanath Mukherjee, Dipak Ghosal,Wireless sensor network survey Department of Computer Science, University of California, Davis, CA 95616, United States

[4] H. W. Kim, H. S. Seo, (2010) "Modeling of Energy-efficient Applicable Routing Algorithm in WSN", International Journal of Digital Content Technology and its Applications, vol. 4, no. 5,pp.13-22.

[5] S.Taruna, Rekha Kumawat, G.N.Purohit , (August 2012) "Multi-Hop Clustering Protocol using Gate wayNodes in Wireless Sensor Networks", International Journal of Wireless &Mobile Networks (IJWMN) VOL.4, No.4.

[6] K. Ramesh and Dr. K. Somasundaram,(November 2011) "Acomparative Study of Cluster head Selection Algorithms in Wireless Sensor Networks", International Journal of Computer Science &Engineering Survey(IJCSES), VOL.2, No.4.

[7] S. Olariu and I. Stojmenovi_c , (2006) "Design Guidelines for MaximizingLifetime and Avoiding Energy Holes in Sensor Networks withUniform Distribution and Uniform Reporting,", Proc. IEEEINFOCOM.

[8] Fengyuan Ren, Jiao Zhang, Tao He, CXhuang Lin and Sajal K. Das, (December 2011)" EBRP: Energy Balanced Routing Protocol for Data gathering in WSNs", IEEE Transactions on parallel and distributed systems, VOL.22, No.12.

[9] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, (2002) "Anapplication-specific protocol architecture for wirelessmicrosensor networks," IEEE Transaction on WirelessCommunications, , vol. 1, no. 4, pp. 660–670.

[10] Fan, X.; Song, Y. October(2007) "Improvement on LEACH Protocol of Wireless Sensor Network",In Proceedings of International Conference on Sensor Technologies and Applications, Valencia, Spain, 14–20, pp. 260–264.

[11] V. Loscrì, G. Morabito and S. Marano. (sept 2005) "A Two-Levels Hierarchy for Low-Energy Adaptive Clustering Hierarchy",Vehicular Technology Conference, VTC 2005 IEEE 62nd, , vol. 3, pp 1809-1813, DOI 10.1109/VETECF 2005 15558418.

[12] Mo Xiaoyan, (2006)" Study and Design on Cluster RoutingProtocols of Wireless Sensor Networks", Ph.D. Dissertation. Zhejiang University, Hangzhou, China,

[13] Ossama Younis, Sonia Fahmy,(October-December 2004)"HEED:Ahybrid,Energy Efficient,Distributed Clustering a proach for Ad Hoc Sensor Networks" IEEE transactions on Mobile Computing Vol .3,No.4, pp 366-379.

[14] Vevik Katiyar, Narottan Chand, (March 2011)"Improvement in LEACH Protocol for large –scale WSNs", Proceedings of (ICETECT), Tamil Nadu, pp1070-1075.

[15] Li-Qing Gue, Yi Xie, Chen-Hui Yang, Zhang-Wei Jing, (11-14 July 2010) "Improvement on LEACH by Combining Adaptive Cluster Head Election and Two-Hop Transmission", Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qindao,

INTENTIONAL BLANK

AN ANALYSIS OF SECURITY CHALLENGES IN MOBILE AD HOC NETWORKS

Ali Dorri and Seyed Reza Kamel and Esmail kheyrkhah

Department of Computer Engineering, Mashhad branch,
Islamic Azad University, Mashhad, Iran.
Alidorri@mshdiau.ac.ir
rezakamel@computer.org
e.kheirkhah@mshdiau.ac.ir

ABSTRACT

Mobile Ad Hoc Network (MANET) is a collection of wireless mobile nodes with restricted transmission range and resources, no fixed infrastructure and quick and easy setup. Because of special characteristics, wide-spread deployment of MANET faced lots of challenges like security, routing and clustering. The security challenges arise due to MANETs self-configuration and self-maintenance capabilities. In this paper, we present an elaborate view of issues in MANET security. We discussed both security services and attacks in detail. Three important parameters in MANET security are defined. Each attack has been analyses briefly based on its own characteristics and behaviour. In addition, defeating approaches against attacks have been evaluated in some important metrics. After analyses and evaluations, future scopes of work have been presented.

KEYWORDS

Mobile Ad Hoc Network (MANET), Security, Attacks on MANET, Security services, Survey.

1. INTRODUCTION

In these years, progresses of wireless technology and increasing popularity of wireless devices, made wireless networks so popular. Mobile Ad Hoc Network (MANET) is an infrastructure-independent network with wireless mobile nodes. MANET is a kind of Ad Hoc networks with special characteristics like open network boundary, dynamic topology, distributed network, fast and quick implementation and hop by hop communications. These characteristics of MANET made it popular, especially in military and disaster management applications. Besides providing benefits, MANET features made it challengeable. Peer to peer applications [1], integration with internet [2], security [3], maintaining network topology [4] and energy [5, 6] are some of the most important challenges in MANET. We briefly discussed MANET challenges in our previous work [7].

In MANET all nodes are free to join and leave the network, also called open network boundary. All intermediate nodes between a source and destination take part in routing, also called hop by hop communications. As communication media is wireless, each node will receive packets in its wireless range, either it has been packets destination or not. These characteristics of MANET increase its vulnerability against malicious behaviours. Therefore, security became the most important challenge in MANET [8].

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 13–25, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41202

In this paper our goal is to provide a comprehensive review on MANET security. Important parameters in security are introduced. Security is divided in two aspects: security services and attacks. A comprehensive review in both security aspects is presented. In addition, we analyses each attack's behaviour in different parameters. Defeating approaches are evaluated in important metrics. Finally, future scopes of work are presented. Rest of this paper is organized as follow: in section 2, three important security parameters in MANET are presented. Section 3 presents two important aspects of security with a discussion on their strategies. Three combinational challenges with security are presented in section 4. Section 5 presents our analyses and classifications on security of MANET and presents some research interest in security. Section 6 introduces open research issues and directions of researches in MANET security. Finally section 7 concludes the paper and introduces best ways to secure MANET and presents some future works.

2. IMPORTANT PARAMETERS IN MANET SECURITY

Because of MANET's special characteristics, there are some important metrics in MANET security that are important in all security approaches; we call them "security parameters". These parameters defined due to MANET special characteristics. Being unaware of these parameters may cause a security approach useless in MANET. Figure 1 shows the relation between security parameters and security challenges. Each security approach must be aware of security parameters as shown in Figure 1. All mechanisms proposed for security aspects, must be aware of these parameters and don't disregard them, otherwise they may be useless in MANET. Security parameters in MANET are as follows:

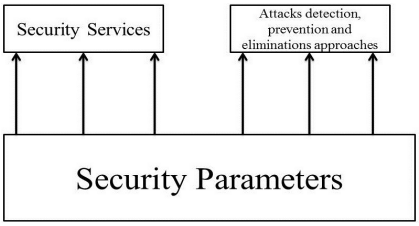


Figure 1.Relation between Security Parameters and Security aspects

Network overhead: Each security mechanism uses a number of control packets. This parameter refers to number of control packets generated by security mechanisms. As MANET uses wireless communications, increasing network overhead may increase collision, congestion and packet loss that may increase packet retransmission.

Processing time: This parameter refers to delay caused by security approach. Because of MANET dynamic topology, neighbours of each node may change in a period of time. As a result, previous information will not be efficient any more.

Energy consumption: Nodes have limited energy resources. One of the major issues is the limited energy, which is usually supplied by node batteries that network nodes possess. Therefore security approach must have low energy consumption.

Each security protocol must be aware of these three important parameters. In some situations a trade-off between these parameters is provided in order to perform a satisfaction level in all of them. Security protocols that disregard these parameters aren't efficient as they waste network resources.

3. MANET SECURITY CHALLENGES

One of the earliest researches in security in MANET was presented in 2002 [9]. Some security challenges in MANET were inherited from ad hoc networks that were research interests since 1999 [10, 11]. Generally there are two important aspects in security: Security services and Attacks. Services refer to some protecting policies in order to make a secure network, while attacks use network vulnerabilities to defeat a security service. In the next two parts, we will discuss these two important aspects of MANET security.

3.1. Security Services

Security services are used to secure a network. Because of special features of MANET, providing these services is a great challenge. For securing MANET a trade-off between these services must be provided, which means if one service guarantees without noticing other services, security system will fail. Providing a trade-off between these security services is depended on network application, but the problem is to provide services one by one in MANET and presenting a way to guarantee each service. We discuss five important security services and their challenges as follows:

Availability: According to this service, each authorized node must have access to all data and services in the network. Availability challenges arise due to MANETs dynamic topology and open boundary. To provide this service, each authorized node must have access to data and services when it needs them. Time spent for accessing is important as time is one of security parameters. By using lots of security and authentication levels, this service is disregarded as passing security levels needs time. Existing approaches for providing this service aren't efficient in MANET and by using them, high degree of distribution, autonomy and dynamicity of MANET will be lost [12]. Authors in [12] provided a new way to solve this problem by using a new trust based clustering approach. In the proposed approach which is called ABTMC (Availability Based Trust Model of Clusters), by using availability based trust model, hostile nodes are identified quickly and should be isolated from the network in a period of time, therefore availability of MANET will be guaranteed.

Authentication: The goal of this service is to provide trustable communications between two different nodes. When a node receives packets from a source, it must be sure about identity of the source node. One way to provide this service is using certifications, whoever in absence of central control unit, key distribution and key management are challengeable. In [13] the authors presented a new way based on trust model and clustering to public the certificate keys. In this case, the network is divided into some clusters and in this clusters public key distribution will be safe by mechanisms provided in the paper. Their simulation results show that, the presented approach is better than PGP. But it has some limitations like clustering. MANET dynamic topology and unpredictable nodes position, made clustering challengeable.

Data confidentially: According to this service, each node or application must have access to specified services that it has the permission to access. Most of services that are provided by data confidentially use encryption methods but in MANET as there is no central management, key distribution faced lots of challenges and in some cases impossible. Authors in [14] proposed a new scheme for reliable data delivery to enhance the data confidentially. The basic idea is to transform a secret message into multiple shares by secret sharing schemes and then deliver the shares via multiple independent paths to the destination. Therefore, even if a small number of nodes that are used to relay the message shares, been compromised, the secret message as a whole is not compromised. Also this way protects network, but packet delivery delay will be increased.

The reason is that, data is sent in multipath. Using multipath delivering causes the variation of delay in packet delivery for different packets. It also leads to out-of-order packet delivery.

Integrity: According to integrity, just authorized nodes can create, edit or delete packets. As an example, Man-In-The-Middle attack is against this service. In this attack, the attacker captures all packets and then removes or modifies them. Authors in [15] presented a mechanism to modify the DSR routing protocol and gain to data integrity by securing the discovering phase of routing protocol.

Non-Repudiation: By using this service, neither source nor destination can repudiate thier behaviour or data. In other words, if a node receives a packet from node 2, and sends a reply, node 2 cannot repudiate the packet that it has been sent. Authors in [16] presented a new approach that is based on grouping and limiting hops in broadcast packets. All group members have a private key to ensure that another node couldn't create packets with its properties. But creating groups in MANET is challengeable.

Till now we discussed security services challenges in MANET. Detecting and eliminating malicious nodes, is another aspect of the MANET security. In the next section, we will discuss some important attacks in MANET and ways to detect and eliminate them.

3.2. Attacks

As described before, MANET has lots of characteristics like hop by hop communications, wireless media, open border and easy to setup, that made it popular. whoever these characteristics made it popular for malicious nodes. Here are some important attacks in MANET:

Black Hole Attack: In this attack, malicious node injects fault routing information to the network and leads packets toward itself, then discards all of them [17-19]. We presented a comprehensive review in this attack in our previous work [20]. We present some ways to detect and eliminate black hole nodes, also we present some classifications in black hole. In [21] authors presented a new approach to detect black hole by sending packets in shared paths and also fetching sequence number of packets. This attack is against routing and is depended on routing protocol.

Worm Hole Attack: In worm hole attack, an attacker records packets at one location of network and tunnels them to another location [22]. Fault routing information could disrupt routes in network [23]. Authors in [24] presented a way to secure MANET against this attack by using encryption and node location information. But as mentioned before, key distribution is a challenge in MANET.

Byzantine attack: In this attack, malicious node injects fault routing information to network, in order to locate packets into a loop [25, 26]. One way to protect network against this attack is using authentication. Authors in [27] presented a mechanism to defeat against this attack using RSA authentication.

Snooping attack: The goal of this attack is accessing to other nodes packets without permission [28]. As in MANET packets transmitted hop by hop, any malicious node can capture others packets.

Routing attack: In this attack, malicious nodes try to modify or delete node's routing tables [17, 18, 29]. Using this attack, malicious nodes destroy routing information. Therefore, packet overhead and processing time will increase.

Resource consumption attack: The goal of this attack is wasting network or node's resources [30, 31]. As a way to do this, malicious nodes lead packets to a loop. Therefore, nodes resources consume so fast.

Session hijacking: Session hijacking is a critical error and gives an opportunity to the malicious node to behave as a legitimate system [32, 33]. Using this attack, malicious node reacts instead of true node in communications. Cryptography is one of the most efficient ways to defeat this attack.

Denial of service: In this attack, malicious node prevents other authorized nodes to access network data or services [34-38]. Using this attack, a specific node or service will be inaccessible and network resources like bandwidth will be wasted. In addition, packet delay and congestion increases.

Jamming attack: Jamming attack is a kind of DOS attack [39]. The objective of a jammer is to interfere with legitimate wireless communications. A jammer can achieve this goal by either preventing a real traffic source from sending out a packet, or by preventing the reception of legitimate packets [40].

Impersonation Attack: Using this attack, attacker can pretend itself as another node and injects fault information to the network [41-43]. As MANET has open border and hop-by-hop communications, it's hardly vulnerable against this attack. In some cases even using authentication is useless.

Modification Attack: In this attack, malicious nodes sniff the network for a period of time. Then, explore wireless frequency and use it to modify packets [44, 45]. Man-in-the-middle is a kind of Modification attack.

Fabrication Attack: In fabrication attack, malicious node destroys routing table of nodes by injecting fault information [46-48]. Malicious node creates fault routing paths. As a result, nodes send their packets in fault routes. Therefore, network resources wasted, packet delivery rate decreased and packet lost will growth.

Man-in-the-middle attack: In this attack, malicious node puts itself between source and destination. Then, captures all packets and drops or modifies them [49-51]. Hop by hop communications are made MANET vulnerable against this attack. Authentication and cryptography are the most effective ways to defeat this attack.

Gray Hole Attack: This attack is similar to black hole. In black hole, malicious node drops all packets, while in this attack, malicious node drops packets with different probabilities [52-55]. As it relays some packets, detecting this attack is more complicated than black hole and some detection approaches like sniffing or watchdog will be useless in it.

Traffic Analyse Attack: The goal of this attack is sniffing network traffic to use them in another attack or in a specific time [44, 56]. Malicious node captures all packets to use them later.

4. INCORPORATING SECURITY AND OTHER CHALLENGES

One way to provide security in MANET, besides decreasing network overhead, is to incorporate security approaches with other challenges. In this way, both challenges are solved by improving security parameters in total. We discuss these combinational approaches as follows:

Secure routing protocols: Using this method, security will be guaranteed in routing phase. When a node wants to create a path to a destination, it uses some mechanisms to select a secure path. Then malicious nodes will be detected and eliminated. Authors in [57] presented a secure routing protocol based on using IPSEC in MANET routing protocols. Authors in [58] presented a trust based security routing protocol to create secure path. In MANET, there is more than one path between two different nodes. Selecting best path based on both routing and security, will improve security parameters.

Security in QOS: Providing security has negative impact on QOS. Therefore, providing QOS beside security is important. Authors in [59] presented a game theory to make a trade-off between security and QOS. Authors in [60] provided an approach that creates QOS aware multipath between source and destination with link information. By providing security in QOS, a level of security and QOS will be guaranteed with low time or network overhead.

Cluster-based Security: These mechanisms use clustering in network to provide more efficient situations for security protocols. Generally in these methods, clusters are used for key distribution or as central management. Using clustering for security goals is important as it could solve problems in key distribution or key management. Authors in [61, 62] provides a key distribution mechanism by using clusters. Using clusters make it easy for security mechanisms, but grouping nodes in clusters and maintaining clusters, are challengeable.

As another benefit of these combinational approaches, there should be a trade-off between challenges based on each one's importance. Based on application, one challenge may be more important than the other one. Then, an algorithm may be proposed with better security parameters. For example, when selecting best path in routing is not as important as security, an approach can choose more secure paths without emphasizing on best routing path. As an example, in AODV routing protocol, the path with low sequence number is chosen as the best path. The reason is that, if there be a malicious node, it will send high sequence number.

5. ANALYSES AND DISCUSSION

Previous sections discussed attacks and security services in MANET. This section presents an analytics and classification on previous issues. In order to analyses attacks and their behaviour, we presented analyses in each attack in Table 1. For each attack five important parameters has been discussed. These parameters are as follows:

- Violated Service: Each attack breaks a security service. We presented the most important defeated service in this column.

- The Proposed Solutions: Some of the most effective approaches to detect and eliminate malicious nodes.

- MANET features which lead to this attack: Each malicious node uses a feature or features of MANET to break the security.

- Attack Type: Lots of researches classified attacks in two mainly class that are as follows: Active attack, Passive attack. In passive attacks, malicious node listens to transmissions without any active injection or effect on network[63]. While, in active attacks malicious node inject information.

- Attack Goal: The most important goal of each attack.

From Table 1it's obvious that lots of attacks are against availability. In availability aimed attacks, malicious nodes inject fault routing information or destroy nodes routing tables in order to defeat availability. As another point in Table 1, sniffing is one of the widely used mechanisms to defeat against attacks. Generally in sniffing mechanisms, sniffer put it-self in promiscuous mode and listen to the network traffic. In this way, it can detect misbehaviour of malicious nodes. As another effective defeating approach, we can name encryption and route information. Encryption mechanism defeat packets against accessing or modification. Malicious nodes can't modify encrypted packets. Whoever, they can drop packets and break availability. Routing information in MANET is nightly variable. Routing information approach, secure MANET using additional controller packets. There are six different defeating mechanisms as presented in Table 1. Defeating approaches are important in MANETs security. Therefore, an analytics on these approaches is presented in Table 2.

Attack Name	Type		Goal				Violated Service	MANET features which lead to this attack	The proposed Solutions
	Active	Passive	Resource consumption	Accessing packets	Modification packets	Drop packets			
Black Hole [17-21]	✓					✓	Availability	Distributed network	Routing information, Sniffing
Worm Hole [22-24]	✓		✓				Availability	Distributed network	Routing information, Encryption, Sniffing
Byzantine [25-27]	✓		✓				Availability	Distributed network	Encryption, Redundancy
Snooping [28]		✓		✓			Data Confidentiality, integrity	Non-centralized	Routing information
Routing [17,18],[29]	✓			✓			Availability	Hop-by-hop communications	Routing information, Authentication
Resource consumption [30,31]	✓		✓				----------	Non-centralized	Encryption, Sniffing
Session hijacking [32-34]		✓		✓			Data Confidentiality	Non-centralized Distributed network	Encryption, Authentication
Denial of service [35-38]	✓					✓	Availability	Non-centralized	Sniffing, Routing information
Jamming [39,40]	✓					✓	Availability	Wireless media	Sniffing, Dynamic frequency
Impersonation [41-43]		✓	✓	✓			Data Confidentiality, Non-repudiation	Open network boundary	Authentication
Modification [44,45]		✓			✓		integrity	Hop-by-hop communication	Encryption
Fabrication [46-48]		✓	✓		✓		Availability	Distributed network	Encryption, Sniffing
Man-in-the-middle [49-52]		✓		✓	✓	✓	Data Confidentiality, integrity	Hop-by-hop communication	Encryption, Authentication
Grey Hole [53-55]	✓					✓	Availability	Distributed network	Routing information, Sniffing
Traffic Analysis [44],[56]		✓		✓			Data Confidentiality	Hop-by-hop communication	Encryption, Authentication

Table 1: Analytics on MANET Attacks**.**

Each proposed solution uses some of network resources in order to detect an attack. Table 2 presents five important parameters for security mechanisms and discussed the effect of each proposed solution in parameters. In order to analyses the energy consumption of approaches; we compare energy consumption in each approach toward others. The word 'low' means it consumes lower energy than other approaches. Also, accuracy refers to ability in defeating single or cooperative malicious nodes. In the case of cooperative attacks, malicious nodes work with each other in order to cover their tracks.

Routing information approach generates controller packets and uses them in order to detect malicious nodes. In addition, in some cases nodes must keep additional routing table like DRI [19]. As mentioned, in sniffing each node must put it-self in promiscuous mode and capture all packets transmission in its range. This feature of sniffing wastes nodes energy. Also it increases process and memory overhead. In the case of cooperative malicious nodes, sniffing is useless as malicious nodes may work with each other to proof them-selves as trustable nodes.

Redundancy and dynamic frequency approaches can't detect the malicious nodes. These approaches can only avoid network from an attack. In the case of misbehaviour, these approaches can detect attack, while they are unable to detect the malicious nodes or eliminate them from whole network. In other work, these approaches can just find the path with malicious nodes.
In routing information, control packets transmission increase processing time. In encryption and authentication, key distribution is an important challenge, because of lack of central infrastructure or key distribution center. Therefore, each malicious node can pretend it-self as a trustable node and take part in key distribution. In redundancy, destination must buffer packets in order to get packets in sequence or to compare them with each other. In addition, it increases traffic overhead by sending duplicated packets. That cause increasing in congestion, packet lost and energy consumption.

In addition of five important parameters, Table 2 presents some limitations on each defeating approaches. These limitations come from each approaches characteristics. These limitations decrease performance or efficiency of each approach.

6. FUTURE DIRECTIONS OF RESEARCHES

Until now we briefly discussed the security challenges in MANET and present some analytics in them. In this section we present open research issues.

Routing information approaches are suitable in all types of MANET. In this approach, reducing packet overhead and processing time, beside increasing accuracy is an important challenge. By increasing accuracy, it can detect cooperative malicious nodes. With decreasing processing time of this approach MANETs flexibility will increase.

Sniffing approach is useful in the case of single attacks, as it is unable to detect cooperative nodes. Whoever, it waste nodes energy and it is not suitable in MANET with high speed nodes. Finding a more effective way to calculate the threshold and present effective detection mechanism forasmuch as decreasing time and packet overhead is the open border of research in sniffing approaches. Beside it, detecting cooperative malicious nodes is challengeable. In order to solve this challenge comparing sniffing with other defeating approaches is recommended.

Table 2: Analysis at the proposed solutions

The proposed Solutions	Energy Consumption	Process Overhead	Memory overhead	Packet Overhead	Accuracy	Limitations
Routing Information [19],[25-27],[33]	Low		✓	✓	S,C	Processing Time
Sniffing [22],[38],[56]	High	✓	✓		S	Cooperative nodes
Encryption [49],[51],[61,62]	Normal	✓	✓	✓	S,C	Absence of Centralized Control, Key distribution
Redundancy [39],[60]	High			✓	None	Packet overhead
Authentication [41,42]	Low	✓		✓	S	Absence of Centralized Control, Key distribution
Dynamic Frequency [39,40]	High			✓	None	Frequency knowledge
The proposed Solutions	Energy Consumption	Process Overhead	Memory overhead	Packet Overhead	Accuracy	Limitations
Routing Information [19],[25-27],[33]	Low		✓	✓	S,C	Processing Time
Sniffing [22],[38],[56]	High	✓	✓		S	Cooperative nodes
Encryption [49],[51],[61,62]	Normal	✓	✓	✓	S,C	Absence of Centralized Control, Key distribution
Redundancy [39],[60]	High			✓	None	Packet overhead
Authentication [41,42]	Low	✓		✓	S	Absence of Centralized Control, Key distribution
Dynamic Frequency [39,40]	High			✓	None	Frequency knowledge

MANET is self-organized, self-configurable network without any centralized control. Therefore, encryption and authentication are challengeable. Key distribution and control unit are the most important challenges. One way to over through these challenges is using clustering; therefore, the Cluster Head can act as the key distributer. Because of MANETs dynamic topology, creating and maintain clusters is highly challengeable. Using fuzzy logic [64] or swarm based [65] is highly recommended for this challenge. As another research interest, decreasing processing time and processing overhead of encryption approach can be mentioned.

Redundancy approaches, generate lots of duplicated packets and waste nodes resources. Also it increases congestion and packet lost. Effectively choosing number of duplicated paths, based on risk level, is highly challengeable. Also combining this approach with some other approaches in order to detect malicious nodes is another challengeable issue.

Dynamic frequency is effective in multi-type MANETs. By using this approach in multi-type MANET, each node secures its packets by sending in different frequencies. In addition, breaking one frequency has no effect on others. This is a challenge in this approach.

7. CONCLUSION

Mobile Ad Hoc Network (MANET) is a kind of Ad hoc network with mobile, wireless nodes. In MANET, all nodes are free to join and leave the network. Features of MANET like dynamic topology and distributed network, made it popular and also challengeable. Open network boundary, dynamic topology and hop by hop communications made security the most important challenge in MANET.

In this paper, we presented a comprehensive review on security challenges in MANET. We divided security in two aspects: security services and attacks. We discussed each one in detail; also we introduced security parameters that are important in MANET security. Finally some analyses and classifications in security approaches were presented. As a result of our analyses, securing network in routing path using route table information, and encryption, are two efficient ways to secure MANET. Each of these approaches have some benefits and limitations. As military applications are one of MANET's applications, encryption is important to avoid malicious nodes to access data packets.

REFERENCES

[1] A. Gantes and j. stucky, "A platform on a Mobile Ad hoc Network challenging collaborative gaming," international symposium on collaborative technologies and systems, 2008.

[2] K. U. R. Khan, R. U. Zaman, and A. V. G. Reddy, "Integrating Mobile Ad Hoc Networks and the Internet challenges and a review of strategies," presented at the 3rd International Conference on Communication Systems Software and Middleware and Workshops, COMSWARE, 2008.

[3] M. Suguna and P. Subathra, " Establishment of stable certificate chains for authentication in mobile ad hoc networks," presented at the International Conference on Recent Trends in Information Technology (ICRTIT), 2011.

[4] H. Nishiyama, T. Ngo, N. Ansari, and N. Kato, "On Minimizing the Impact of Mobility on Topology Control in Mobile Ad Hoc Networks," Wireless Communications, IEEE Transactions, 2012.

[5] F. D. Rango, M. Fotino, and S. Marano, "EE-OLSR: Energy Efficient OLSR routing protocol for Mobile ad-hoc Networks," presented at the Military Communications Conference, MILCOM, 2008.

[6] K. Du and Y. Yang, "Policy-Based Time Slot Assignment algorithm in a MANET(PBTSA)," presented at the 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, ASID, 2009.

[7] A. dorri, S. R. kamel, and E. kheirkhah, "ارائـه شـده ی‌و راهکارها MANET ی‌شـــبکه ی‌هابـر چـالش ی‌ل‌ی‌تحـل آن ی‌هابهبـــود عملکـــرد شـــبکه در بـــرابر چـــالش ی‌بـرا," presented at the first national conference on electronical and computer north Iran (Bandar Anzali), 2014.

[8] R. Sheikh, M. S. Chande, and D. K. Mishra, "Security issues in MANET:A review," presented at the Seventh International Conference On Wireless And Optical Communications Networks (WOCN), 2010.

[9] H. Deng, W. Li, and D. P. Agrawal, "Routing security in wireless ad hoc networks,," Communications Magazine,IEEE, 2002.

[10] Y.Z.a and W. Lee, "Intrusion Detection in Wireless Ad-Hoc networks," presented at the 6th Int'l. Conf. Mobile Comp. Net., MobiCom, 2000.

[11] F.S.a and R. Anderson, "The Resurrecting Ducking: Security Issues for Ad-Hoc Wireless Networks," 7th Int'l. Wksp on Security Protocols. Proc., LNC, 1999.

[12] X. Zhao, Z. You, Z. Zhao, D. Chen, and F. Peng, "Availability Based Trust Model of Clusters for MANET," presented at the 7th International Conference on Service Systems and Service Management (ICSSSM), 2011.

[13] E. C.H.Ngai and L. M. R, "Trust and clustering-based Authentication Services in Mobile ad hoc networks," presented at the proceeding of the 24th international conference on Distributed Computing systems Workshops 2004.

[14] W. Lou, W. Liu, and Y. Fang, "SPREAD: enhancing data confidentiality in mobile ad hoc networks," presented at the Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, 2004.

[15] S. Rana and A. Kapil, "Security-Aware Efficient Route Discovery for DSR in MANET," Information and Communication Technologies, Communications in Computer and Information Science, vol. 101, pp. 186-194, 2010.
[16] X. Lv and H. Li, "Secure group communication with both confidentiality and non-repudiation for mobile ad-hoc networks," Information Security, IET, vol. 7, 2013.
[17] S.a.A.k.G, H.o.d.R.m, and S. sharma, "A Comprehensive Review of Security Issues in Manets," International Journal of Computer Applications vol. 69 2013.
[18] V.P. and R. P. Goyal, "MANET: Vulnerabilities, Challenges, Attacks, Application," IJCEM International journal of Computational Engineering & management, vol. 11, 2011.
[19] A. MISHRA, R. Jaiswal, and S. Sharma, " A novel approach for detecting and eliminating cooperative black hole attack using advanced DRI table in Ad hoc Network," presented at the 3rd International Conference on Advance Computing Conference (IACC), 2013
[20] A. dorri and T. m. k. zade, "کشـــف آن یهااه و راهیس یحفـره یل حملـهیتحـل," presented at the first regional conference on optimizing and soft computing in electronic and computer engineering, 2014.
[21] N. S. A. Sharma, "The Black-Hole Node Attack in MANET," presented at the Second International Conference on Advanced Computing & Communication Technologies, ACCT '12 Proceedings of the, 2012.
[22] M. A. Gorlatova, P. C. Mason, M. Wang, and L. Lamont, " Detecting Wormhole Attacks in Mobile Ad Hoc Networks through Protocol Breaking and Packet Timing Analysis," Military Communications Conference, IEEE, MILCOM, 2006.
[23] S. Keer and A. Suryavanshi, "To prevent wormhole attacks using wireless protocol in MANET," presented at the nternational Conference on Computer and Communication Technology (ICCCT), 2010.
[24] Z. A. Khan and M. H. Islam, "Wormhole attack: A new detection technique," presented at the international conference on Emerging Technologies (ICET), 2012.
[25] M. Yu, M. C. Zhou, and W. Su, "A Secure Routing Protocol Against Byzantine Attacks for MANETs in Adversarial Environments," IEEE Transactions on Vehicular Technology, vol. 58
[26] G. Singla, M. S. Sathisha, A. Ranjan, S. D., and P. Kumara, "Implementation of protected routing to defend byzantine attacks for MANET's," International Journal of Advanced Research in Computer Science, vol. 3, p. 109, 2012.
[27] G. Singla and P. Kaliyar, "A Secure Routing Protocol for MANETs Against Byzantine Attacks," Computer Networks & Communications (NetCom), Lecture Notes in Electrical Engineering, vol. 131, pp. 571-578, 2013.
[28] S. Shaw, K. Orea, P. Venkateswaran, and R. Nandi, " Simulation and Performance Analysis of OLSR under Identity Spoofing Attack for Mobile Ad-Hoc Networks," Computer Networks and Information Technologies Communications in Computer and Information Science, vol. 142, pp. 308-310, 2011.
[29] B. Kannhavong, H. Nakayama, Y. Nemoto, and N. Kato, "A survey of routing attacks in mobile ad hoc networks," Wireless Communications, IEEE Transactions, vol. 14
[30] M. Abdelhaq, R. Hassan, and R. Alsaqour, "Using Dendritic Cell Algorithm to Detect the Resource Consumption Attack over MANET," Software Engineering and Computer Systems Communications in Computer and Information Science vol. 181, pp. 429-442, 2011.
[31] L. Rajeswari, A. Prema, R. A. Xavier, and A. Kannan, "Enhanced intrusion detection techniques for mobile ad hoc networks," presented at the International Conference on Information and Communication Technology in Electrical Sciences (ICTES), 2007.
[32] A. K. Rai, R. R. Tewari, and S. K. Upadhyay, "different type of attacks on integrated MANET-internet comunication," international jornal of computer science and security (IJCSS), vol. 4.
[33] J. Y. Kim, H. K. Choi, and S. Song, "A secure and lightweight approach for routing optimization in mobile IPv6," EURASIP Journal on Wireless Communications and Networking - Special issue on wireless network security, vol. 7, 2009.
[34] Supriya and M. Khari, "Mobile Ad Hoc Netwoks Security Attacks and Secured Routing Protocols: A Survey," Advances in Computer Science and Information Technology. Networks and Communications Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 84, pp. 119-124, 2012.
[35] J. Soryal and T. Saadawi, "IEEE 802.11 Denial of Service attack detection in MANET," Wireless Telecommunications Symposium (WTS), 2012.
[36] R. H. Jhaveri, S. J. Patel, and D. C. Jinwala, "DoS Attacks in Mobile Ad Hoc Networks: A Survey," presented at the Second International Conference on Advanced Computing & Communication Technologies (ACCT), 2012

[37] A. Michael and Nadeem, "Adaptive intrusion detection & prevention of denial of service attacks in MANETs," presented at the IWCMC '09 Proceedings of the International Conference on Wireless Communications and Mobile Computing, Connecting the World Wirelessly, 2009.
[38] J. Su and H. Liu, "Protecting Flow Design for DoS Attack and Defense at the MAC Layer in Mobile Ad Hoc Network," Applied Informatics and Communication Communications in Computer and Information Science, vol. 224, pp. 233-240, 2011.
[39] A. Hamieh and J. Ben-othman, "Detection of Jamming Attacks in Wireless Ad Hoc Networks Using Error Distribution," presented at the International Conference on Communications, ICC '09. IEEE, 2009.
[40] J. Ben-othman and A. Hamieh, "Defending method against jamming attack in wireless ad hoc networks," presented at the 34th Conference on Local Computer Networks, LCN, IEEE, 2009.
[41] D. Glynos, P. Kotzanikolaou, and C. Douligeris, "Preventing impersonation attacks in MANET with multi-factor authentication," hird International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WIOPT, 2005.
[42] C. Douligeris, P. Kotzanikolaou, and D. Glynos, "Preventing Impersonation Attacks in MANET with Multi-Factor Authentication," WIOPT '05 Proceedings of the Third International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 2005.
[43] M. Barbeau, J. Hall, and E. Kranakis, "Detecting Impersonation Attacks in Future Wireless and Mobile Networks," Secure Mobile Ad-hoc Networks and Sensors Lecture Notes in Computer Science, vol. 4074, pp. 80-95, 2006.
[44] N. Dixit, S. Agrawal, and V. K. Singh, "A Proposed Solution for security Issues In MANETs," International Journal of Engineering Research & Technology(IJERT), vol. 2, 2013.
[45] Vaithiyanathan, S. R. Gracelin, E. N. Edna, and S. Radha, "A Novel Method for Detection and Elimination of Modification Attack and TTL Attack in NTP Based Routing Algorithm," presented at the International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), 2010
[46] P. Yi, X. Jiang, and Y. Wu, "Distributed intrusion detection for mobile ad hoc networks," Journal on Systems Engineering and Electronics, IEEE, vol. 19, 2008.
[47] S. R. Afzal, S. Biswas, J. B. Koh, T. Raza, and m. authors, "RSRP: A Robust Secure Routing Protocol for Mobile Ad Hoc Networks," presented at the Wireless Communications and Networking Conference,WCNC, IEEE, 2008.
[48] P. T. Tharani, K. Muthupriya, and C. Timotta, "Secured consistent network for coping up with gabrication attack in MANET," international journal of Emerging Technology and Advanced Engeineering, vol. 3, 2013.
[49] D. Sharma, P. G. Shah, and X. Huang, "Protecting from Attacking the Man-in-Middle in Wireless Sensor Networks with Elliptic Curve Cryptography Key Exchange," presented at the NSS '10 Proceedings of the Fourth International Conference on Network and System Security, 2010.
[50] K. Vishnu, "A new kind of transport layer attack in wireless Ad Hoc Networks," presented at the International Conference on Wireless Communications, Networking and Information Security (WCNIS), 2010
[51] X. Zou, A. Thukral, and B. Ramamurthy, "An Authenticated Key Agreement Protocol for Mobile Ad Hoc Networks," Mobile Ad-hoc and Sensor Networks Lecture Notes in Computer Science, vol. 4325, pp. 509-520, 2006.
[52] J. Liu, F. Fu, J. Xiao, and Y. Lu, "Secure Routing for Mobile Ad Hoc Networks," presented at the Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD, 2007.
[53] J. Sen, B. Tata, M. Chandra, S. Harihara, and H. Reddy, "A mechanism for detection of gray hole attack in mobile Ad Hoc networks," presented at the 6th International Conference on Information, Communications & Signal Processing, 2007
[54] G. Usha and S. Bose, "Impact of Gray hole attack on adhoc networks," presented at the International Conference on Information Communication and Embedded Systems (ICICES), 2013
[55] G. Xiaopeng and C. Wei, "A Novel Gray Hole Attack Detection Scheme for Mobile Ad-Hoc Networks," presented at the IFIP International Conference on Network and Parallel Computing Workshops, NPC Workshops, 2007.
[56] C. Gray, J. Byrnes, and S. Nelakuditi, "Pair-wise Resistence to traffic Analysis in MANETs," ACM SIGMOBILE Mobile Computing and Communications Review, 2008.
[57] E. A. Panaousis, T. A. Ramrekha, and C. Politis, "Secure routing for supporting ad-hoc extreme emergency infrastructures," Future Network and Mobile Summit, 2010.

[58] M. Salmanian and M. Li, "Enabling secure and reliable policy-based routing in MANETs," presented at the military communications conference, MILCOM, 2012.
[59] F. R. Yu, H. Tang, S. Bu, and D. Zheng, "Security and quality of service (QoS) co-design in cooperative mobile ad hocnetworks," EURASIP Journal on Wireless Communications and Networking - Special issue on wireless network security, 2013.
[60] R. Gujral, A. Kapil, and " ", Volume , , pp "Secure QoS Enabled On-Demand Link-State Multipath Routing in MANETs," Information Processing and Management Communications in Computer and Information Science, vol. 70, pp. 250-257, 2010.
[61] A. El-Sayed, "Clustering Based Group Key Management for MANET," Advances in Security of Information and Communication, Networks Communications in Computer and Information Science, vol. 381, pp. 11-26, 2013.
[62] M. S. Zefreh, A. Fanian, S. M. Sajadieh, P.Khadivi, and M. Berenjkoub, "A Cluster-Based Key Establishment Protocol for Wireless Mobile Ad Hoc Networks," Advances in Computer Science and Engineering Communications in Computer and Information Science, vol. 6, pp. 585-592, 2009.
[63] L. Yingbin, H. V. Poor, and Y. Lei, "Secrecy Throughput of MANETs Under Passive and Active Attacks," Information Theory, IEEE Transactions on, vol. 57, pp. 6692-6702, 2011.
[64] W. El-Hajj, D. Kountanis, A. Al-Fuqaha, and M. Guizani, "A Fuzzy-Based Hierarchical Energy Efficient Routing Protocol for Large Scale Mobile Ad Hoc Networks (FEER)," in Communications, 2006. ICC '06. IEEE International Conference on, 2006, pp. 3585-3590.
[65] M. Achankunju, R. Pushpalakshmi, and A. A. Kumar, "Particle swarm optimization based secure QoS clustering for mobile ad hoc network," in Communications and Signal Processing (ICCSP), 2013 International Conference on, 2013, pp. 315-320.

AUTHORS

Ali Dorri received his B.S. degree in computer engineering from Bojnord University, Iran, in 2012, and now is student in M.S in software engineering in Mashhad branch, Islamic Azad University, Mashhad, Iran. His research interests cover Wireless Sensor Networks (WSN), Mobile Ad hoc Network (MANET) and specially Security challenges.

Dr. Seyed Reza Kamel Tabbakh is with the Department of Software Engineering, Faculty of Engineering, Islamic Azad University - Mashhad branch, Mashhad, Iran. He received his PhD in communication and network engineering from University Putra Malaysia (UPM) in 2011. He received his BSc and MSc in software engineering from Islamic Azad University, Mashhad branch and Islamic Azad University, South Tehran branch, Iran in 1999 and 2001 respectively. His research interests include IPv6 networks, routing and security. During his studies, he has published several papers in International journals and conferences.

Esmaeil Kheirkhah received his Bachelor and Master in Computer Science and Mathematics from Islamic Azad University, Mashhad, Iran in 1992 and 1996 respectively. He also received his PhD in Computer Science from National University of Malaysia (UKM) in 2010. He is currently an assistant professor at the Islamic Azad University of Mashhad. His research interests include the Software Engineering, Requirements Engineering, End-User Computing, and semantic-enabled software engineering.

INTENTIONAL BLANK

A FUZZY-BASED CONGESTION CONTROLLER FOR CONTROL AND BALANCE CONGESTION IN GRIED-BASED WSN

Ali Dorri and Seyed Reza Kamel

Department of Computer Engineering, Mashhad Branch,
Islamic Azad University, Mashhad, Iran
Alidorri@mshdiau.ac.ir
Rezakamel@computer.org

ABSTRACT

A Wireless Sensor Network (WSN) is deployed with a large number of sensors with limited power supply in a wide geographically area. These sensors collect information depending on application. The sensors transmit the data towards a base station called sink. Due to the relatively high node density and source-to-sink communication pattern, congestion is a critical issue in WSN. Congestion not only causes packet loss, but also leads to excessive energy consumption as well as delay. To address this problem, in this paper we propose a new fuzzy logic based mechanism to detect and control congestion in WSN. In the proposed approach, a Monitor Node for each grid in congestion candidate region performs a fuzzy control to avoid increasing congestion. Fuzzy controller's inputs are continually fetched from the network by the Monitor Node. Simulation results show that our approach has higher packet delivery ratio and lower packet loss than existing approaches.

KEYWORDS

Fuzzy Logic, Wireless Sensor Network, Congestion Control, Packet Lost.

1. INTRODUCTION

A Wireless Sensor Network (WSN) consists of spatially distributed autonomous wireless sensor nodes to cooperatively monitor physical or environmental conditions, such as temperature, sound and pressure. In addition, WSN is a network made of hundreds or thousands of sensor nodes which are densely deployed in hazardous/unattended environment with capability of sensing, computing and sending information wirelessly to the base station (also called sink) via neighbour nodes. Figure 1 shows a WSN that collect information and send it to the sink.

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 27–36, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41203

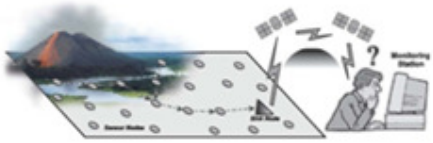


Figure 1. Wireless Sensor Network [1]

In WSN special applications, once an event occurs, a sudden surge of data traffic will be triggered by all sensor nodes in the event area, which may easily lead to congestion. Congestion in WSN has a negative impact on network performance. It increases the packet loss, end-to-end delay and wastes nodes energy. When a packet is lost, source node must retransmit it again. Therefore, node's energy is wasted and network lifetime will be decreased [2,3]. However, some characteristics of WSN, such as constrained resources and leak of centralized coordination, make the congestion problem in WSN more challengeable than any other networks. In WSN all nodes send their sensed data to the sink. This flow of packet (also called source-to-sink traffic), increase congestion probability and energy consumption in nodes near the sink. The reason is that, the neighbouring nodes of the sink should forward other nodes packets in addition to their own traffic [4,5].

To address these challenges, we present a fuzzy congestion control. In the proposed approach, the network is divided into grids by the sink. Then, the sink specifies congestion candidate areas by use of a calculated threshold. In each congestion candidate area, a Monitor Node (MN) uses a fuzzy controller to detect and avoid congestion in its grid. Congestion level of each grid maybe different from other grids and if the congestion level of any congestion candidate area reaches to the acting level, the Border Node forwards packets out of the congested area. Therefore, congestion and packet loss will be decreased. Simulation results show that the presented fuzzy-based system decreases the packet loss in congested area and increases packet delivery ratio to the sink. The remainder of this paper is organized as follows: section 2 presents a literature review of the congestion detection and control mechanisms. Section 3 provides detailed description of the proposed fuzzy controller approach. Performance evaluation of the proposed approach is presented in section 4. Finally section 5 concludes and discusses the future directions of this research.

2. RELATED WORKS

In literature many congestion control schemes have been proposed for WSN. Congestion control schemes for WSN either focus on MAC layer or on both MAC and network layer [6]. Authors in [7] presented an approach based on a threshold. This threshold refers to ratio of received packets to serviced packets. In this approach each node has a priority. To detect congestion, both threshold and priority of nodes are influenced. Proposed approach is efficient in term of Quality of Service (QOS) as it sends data through multipath. Authors in [8] presented a Medium Access Control (MAC) technique to coordinate the access of nodes to the shared medium. It uses the queue buffer length of the sensor nodes to estimate the congestion. Then the traffic dynamically disseminates along with classifying nodes into different priority classes to provide a congestion-free routing path to the destination with improved QOS. In addition, it uses multiple forwarder traffic diffusion, which has advantages like increasing network reliability and reducing congestion. In [9] an optimal routing algorithm that allows optimizing transmission between the peripheral nodes and central node is presented, in order to increase the residual energy of the network. This protocol only aims to provide routing fidelity and does not address time transmission requirements. Authors in [10] presented a cross-layer congestion controller that has three parts: 1) multipath routing 2) adjusted ratio 3) application oriented design. In this algorithm, each node has multiple downstream nodes to be transmitted. The probability of forwarding nodes

and the rate of sending packets can be dynamically adjusted according to the congestion state. Authors in [11] presented a novel approach based on bird's behaviour. The proposed approach is simple to implement at the individual node, involving minimal information exchange. In addition, it displays global self-properties and emergent behaviour, achieved collectively without explicitly programming these properties into individual packets. Performance evaluations show the effectiveness of the proposed Flock-based Congestion Control (Flock-CC) mechanism in dynamically balancing the offered load by effectively exploiting available network resources and moving packets to the sink. Furthermore, Flock-CC provides graceful performance degradation in terms of packet delivery ratio, packet loss, delay and energy consumption under low, high and extreme traffic loads. In addition, the proposed approach achieves robustness against failure and also has scalability in different network sizes and outperforms typical conventional approaches.

3. THE PROPOSED APPROACH

In previous section, we presented a literature review on congestion controller mechanisms in WSN. In this section, we briefly discuss our fuzzy-based congestion controller. Fuzzy system is used in order to increase accuracy of the congestion controller system. Fuzzy controllers convert crisp inputs to fuzzy based inputs, then by using a rule base, fuzzy system determines an action as the main output. The basic idea of a fuzzy controller is presented in Figure 2.

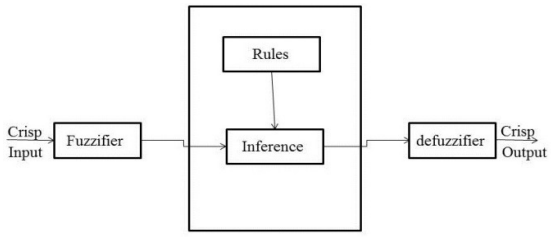


Figure 2. Fuzzy Controller System Functionality

In presented approach, fuzzy logic controller is considered as the kernel of the algorithm. It is associated with the Monitor Nodes (MNs) to detect congestion, based on information which comes from the network. In the propose approach, MN continuously monitors its grid and fetches fuzzy controller's metrics. Therefore, the fuzzy system calculates congestion level in each grid continuously and dynamically. We describe our fuzzy-based congestion controller in three phases that are as follows:

1) Congestion candidate generation
2) Congestion identification
3) Generating new phase

We discuss each phase briefly in the rest of this section.

3.1. Congestion Candidate Generation

After the sensors establishment, the sink uses nodes geographically information in order to divide network to some equal grids. In some applications of WSN, sensors deployed randomly in order to monitor environment and as the grids are equal in area, number of sensors in each grid maybe different from other grids. After dividing the network into grids, sink allocates an ID to each grid called Grid_ID. Then, it sends the Grid_ID of each grid to its members. The next step is choosing the Monitor Node (MN). Responsibility of the MN is to monitor its own grid continuously in order to detect congestion. Sink choose the sensor with highest reminded energy as the MN. In

the case that two sensors have the same reminded energy, the nearest sensor to the sink is chosen as the MN. The MN needs to perform some monitoring works that consumes energy. Therefore, the sensor with highest reminded energy is chosen as the MN.

The presented fuzzy system uses each grid's density in order to detect congestion areas. Therefore, the first duty of each MN is to calculate its grids density. Whoever, at the beginning, sink knows the number of sensors in each grid, but in the case that a sensor dies as lack of energy, the MN must recalculate its grids density. When sensors energy reaches to the alarm level, it sends a packet for MN and aware MN of its death. The alarm levels energy is enough just for sending a packet to MN.

In order to calculate grids density, the MN put Grid_ID in a packet and broadcast it for its neighbours. Each sensor that receives the packet compares its own Grid_ID with the received Grid_ID. If both are the same, the sensor sends a replay to the MN and then rebroadcasts packet. Sensor drops the packet, either if it received the same packet before or if packet's Grid_ID is not the same as node's Grid_ID.

Each MN sends the density of its grid to the sink. After receiving all grids densities, sink calculates the average of densities and uses it as a threshold for defining congested areas. Each grid with higher density than the calculated threshold is marked as the congested candidate area. If any change in number of sensors happened, the MN sends its new density for the sink. Sink recalculates the new threshold and updates congested candidate regions. High density regions have higher congestion probability and higher collision and packet lost rate. A summary of this phase is shown in Figure 3.

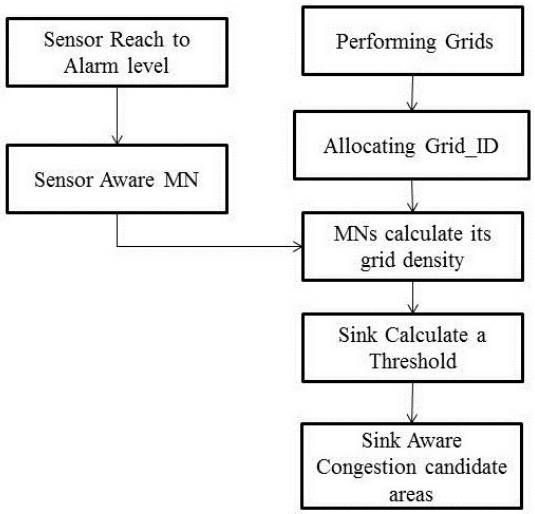


Figure 3. Steps for Congested Region discovery

3.2. Congestion Identification

In each congestion candidate grid, the MN has responsibility for detecting the congestion. In each MN there is a fuzzy-based system, which controls the congestion in grid and balance traffic in order to reduce the congestion. Fuzzy system uses three different metrics in order to decrease congestion and reduce its effects. Fuzzy controller for each MN is shown in Figure 4. MN continuously calculates three input parameters and then uses fuzzy logic to determine specific

level of congestion. Based on the congestion level, either packets forward through the grid or through the relay nodes.

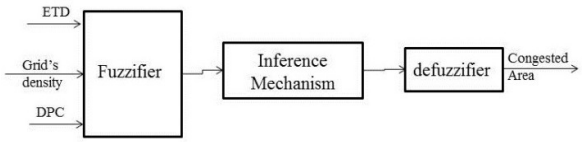


Figure 4. The Proposed Fuzzy Logic Controller

Three fuzzy input parameters are as follows:

ETD (Extended Transmission Delay): The ETD metric is the transition time that is required to transmit a packet to the next hop. The metric is calculated using the convex combination presented in Formula.1.

$$ETD(t)=\alpha * Delay(t) + (1-\alpha) * ETD(t-1) \qquad \text{Formula.1}$$

In this formula, Delay(t) determines delay in time t, and α is weight value. α is used to determine the priority of delay or previous ETD.

Grid's Density: High density of nodes in each grid can cause more congestion. In addition, grids density maybe increased or decreased during network lifetime. The reason is that a sensor may die or new sensors may add to the network. Number of sensors in each grid has a direct effect on congestion. Therefore, number of sensors is an important metric in the proposed fuzzy system.

DPC (Dropped Packets): Dropped packets, refers to number of lost packets in each grid. Packet loss is the result of congestion. Therefore, increasing packet loss means increasing congestion.

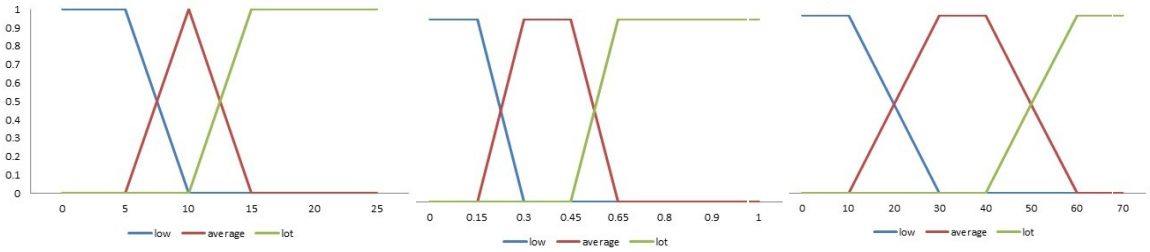
These parameters have membership functions that are presented in Figure 5. By use of these membership functions, a rule base is designed for fuzzy system. When MN calculates inputs, fuzzy system applies them in the rule base. Fuzzy system output determines three different actions that are as follows:

Action1) relay all packets through the grid

Action2) relay half of the packets through the grid and half through the relay nodes

Action3) relay all packets through the relay nodes

We will discuss the relay nodes and these actions in the next phase. Using presented mechanism in this phase, fuzzy controller detects congestion and try to reduce the packets in the congested area.



a) Grids Density b) ETD c) Dropped packets

Figure 5. Membership Functions Diagrams

3.3. Generating New Path

When fuzzy system determines the action, the MN performs the specified action. If there is no congestion, output will be action1. In this case, there is no need for the MN to perform any action. But for other two actions, the MN must informs its grid's Border Node (BN) about the congestion level. At the first time, the MN sends a packet to the sink and asks for its BN's ID. The sink selects the last node in grid as the BN. The detail of selecting BN is presented in [13]. Figure 6 shows the position of BNs and MNs in the network. After selecting the BN, the sink informs the MN of its grids BN. BN has the responsibility of relaying the packets either through the grid or through the relay nodes. The BN relay packets based on action level and this will continue until congestion level in the grid reduce to action1. When the BN gets informed of any changes in congestion level by the MN, it sends packets through relay path based of congestion level. Since radio frequency of relay nodes is different from ordinal sensors, sending packet using relay path has no effect on congestion in the grid. Sink selects the best relay path for BN based on [13].

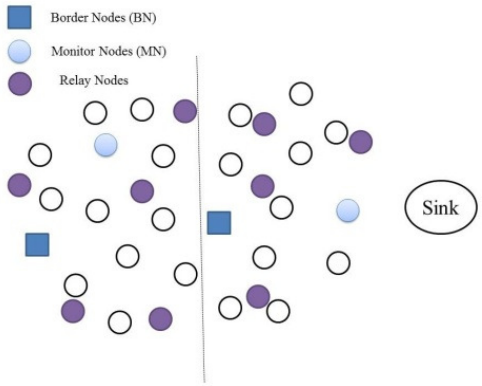


Figure 6. Example of BN and MN Assignments

In this section, we briefly discussed our proposed approach. A summary of the proposed approach is shown in Figure 7.

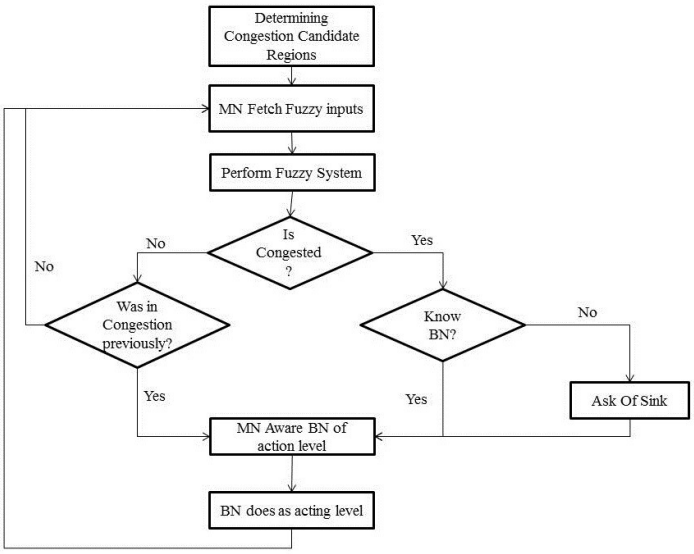


Figure 7. The Proposed Approach

4. SIMULATION RESULTS

To show the advantages of the proposed approach in compare with the Base Work (BW) [12], both approaches implemented using Opnet Modular 14.5 simulator. The simulation was performed using a WSN of size 300 m* 300 m. Table 1, lists simulation parameters used in our study.

Both the BW and our approach were implemented using two different scenarios. For both scenarios two parameters have been measured and evaluated. These parameters are as follows:

Packet delivery: number of packets reached to the sink
Number of dropped packets: number of packet lost because of congestion in congestion candidate grids.

Table 1. Simulation Parameters.

Parameter	Value
Simulation duration	180 sec
Number of nodes	26
Transmission range	20 m
Traffic type	CBR(UDP)
Packet rate	2 packets/sec
Data payload	512 byte/packet
Number of relay nodes	6

The aim of the proposed fuzzy congestion controller is to detect and avoid increasing congestion in congestion candidate regions. When a packet is lost, it should be retransmitted by the source node which wastes node's energy and the network bandwidth.

Figure 8 presents packet delivery rate for both approaches.

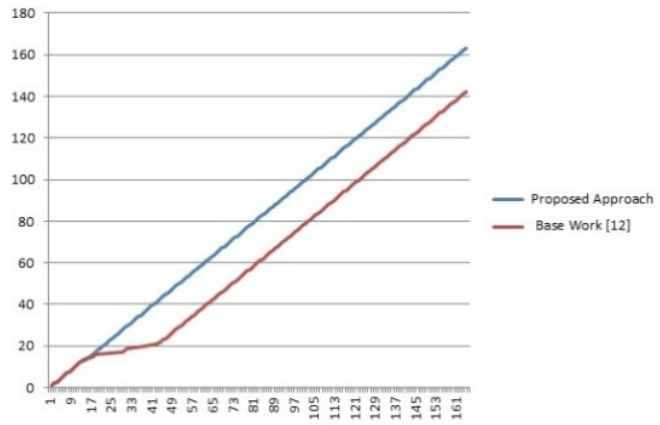


Figure 8. Packet Delivery Rate

In this figure horizontal axis refers to time and the vertical axis, to number of packets delivered to the sink. This figure shows the packet delivery ratio in the sink. Firstly in this figure, both curves have overlap as our fuzzy mechanism hasn't detected congestion yet. As fuzzy controller uses three metrics to detect congestion, after a while, it detects a level of congestion and starts to lead packets toward relay nodes, based on congestion level. As number of packet lost increased, BW starts to send all packets through relay nodes. Therefore, after a while both approach send packets through relay nodes and packet delivery ratio will be the same in the sink for both approaches.

Figure 9 shows the packet lost figure for both approaches.

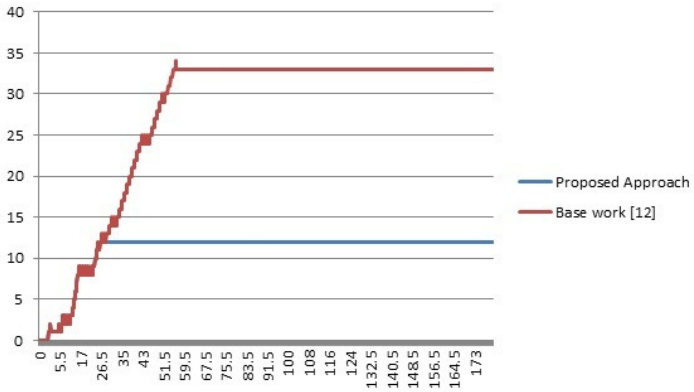


Figure 9. Packet Loss in Congested Grid rate

Vertical axis refers to packet lost in congested grid, and horizontal axis refers to time. Like Figure 8, in this figure, firstly packet lost in both approaches is the same. After increasing packet lost, fuzzy system detects congestion and according to congestion level, it sends packets through either congested grid or relay nodes. As a result, congestion and packet lost in congested candidate grid will be decreased. Whoever, after a while, as both approaches relay packets through the relay nodes, packet lost is equal in both approaches.

Referred to simulation results, proposed approach increases packet delivery ratio in the sink and decreases packet lost in congested grids. In addition, fuzzy system controls the congestion and avoids growth in congestion. Also, fuzzy system is more flexible as it continuously fetches

metrics from the network. By increasing the number of curves in membership functions of input metrics, the accuracy of system increases. As a result, number of actions and control ratio increases. When fuzzy system reaches to acting level, proposed approach controls congestion and helps the grid to balance traffic by leading some packets through relay paths.

5. CONCLUSION AND FUTURE WORK

Wireless Sensor Network (WSN) is a set of sensor nodes, which are distributed in an area. In WSN all sensor nodes sends their packets hop-by-hop to the sink, therefore, nodes nearer to the sink should forward their own packets and other nodes packets. This feature of WSN, increases congestion and packet lost in nodes, especially in nearer nodes to the sink. Using congestion controller mechanisms can avoid or control congestion in WSN. As a result of decreasing congestion, packet loss and energy consumption of the nodes will decrease. In this paper, we proposed a novel approach based on fuzzy logic to control and balance congestion in grid-based WSN. Presented approach uses three parameters that dynamically and continuously fetches from the network in order to detect congestion regions. When fuzzy system detects a level of congestion it relay some packets out of the grid in order to control and balance congestion. As our future work, we decided to make our work more energy aware as energy is the most important parameter in WSN.

REFERENCES

[1] L.A. Villas, A. Boukerche, H.S. Ramos, H.A.B.F. de Oliveira, "DRINA: A Lightweight and Reliable Routing Approach for In-Network Aggregation in Wireless Sensor Networks", IEEE Transactions on Computers, (Volume:62 , Issue: 4), 2013.

[2] R. Annie Uthra, S.V. Kasmir Raja, " Energy Efficient Congestion Control in Wireless Sensor Network", Recent Advances in Intelligent Informatics Advances in Intelligent Systems and Computing Volume 235, pp 331-341, 2014.

[3] Sh. Borasia, V. Raisinghani," A Review of Congestion Control Mechanisms for Wireless Sensor Networks", Technology Systems and Management Communications in Computer and Information Science Volume 145, pp 201-206, 2011.

[4] C. Karakus, A.C. Gurbuz, B. Tavli," Analysis of Energy Efficiency of Compressive Sensing in Wireless Sensor Networks", Sensors Journal, IEEE (Volume:13 , Issue: 5), 2013.

[5] C. Sergiou, V. Vassiliou, " Study of lifetime extension in wireless sensor networks through congestion control algorithms", IEEE Symposium on Computers and Communications (ISCC), 2011.

[6] J. Zhao, L. Wang, S. Li, X. Liu, " A Survey of Congestion Control Mechanisms in Wireless Sensor Networks", in Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010.

[7] C. Wang, B. Li, K. Sohraby, M. Daneshmand, "Upstream congestion control in wireless sensor networks through cross-layer optimization". IEEE Journal on Selected Areas in Communications, (Volume:25 , Issue: 4).

[8] U.S. Visweswaraiya, K.S. Gurumurthy, " A Novel, Dynamic Data Dissemination [D3] Technique for Congestion Avoidance/Control in High Speed Wireless Multimedia Sensor Networks", Fifth International Conference on Computational Intelligence, Modelling and Simulation (CIMSim), 2013.

[9] E. Hajian, K. Jamshidi, A. Bohlooli, "Improve energy efficiency routing in WSN by using automata", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol. 1, No.2, 2010.

[10] Li. Zilong, W. Zou, T. Qi, " A cross-layer congestion control strategy in wireless sensor network", in 4th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT), 2011.

[11] P. Antoniou, A. Pitsillides, T. Blackwell, A. Engelbrecht, L. Michael," Congestion control in wireless sensor networks based on bird flocking behavior", Computer Networks Volume 57, Issue 5, 7 April 2013, Pages 1167–1191.

[12] H. Cha, K. Kim, S.Yoo, " A node placement algorithm for avoiding congestion regions in wireless sensor networks", in Third International Conference on Ubiquitous and Future Networks (ICUFN), 2011.

[13] J. Mena, V. Kalogeraki, " Dynamic Relay Node Placement in Wireless Sensor Networks" in International Symposium on Applications and the Internet, SAINT ,2008.

AUTHORS

Ali Dorri received his B.S. degree in computer engineering from Bojnord University, Iran, in 2012, and now is student in M.S in software engineering in Mashhad branch, Islamic Azad University, Mashhad, Iran. His research interests cover Wireless Sensor Networks (WSN), Mobile Ad hoc Network (MANET) and specially Security challenges.

Dr. Seyed Reza Kamel Tabbakh is with the Department of Software Engineering, Faculty of Engineering, Islamic Azad University - Mashhad branch, Mashhad, Iran. He received his PhD in communication and network engineering from University Putra Malaysia (UPM) in 2011. He received his BSc and MSc in software engineering from Islamic Azad University, Mashhad branch and Islamic Azad University, South Tehran branch, Iran in 1999 and 2001 respectively. His research interests include IPv6 networks, routing and security. During his studies, he has published several papers in International journals and conferences.

THE PROPOSAL OF GIVING TWO RECEIPTS FOR VOTERS TO INCREASE THE SECURITY OF ELECTRONIC VOTING

Abbas Akkasi[1], Ali Khaleghi[2], Mohammad Jafarabad[3], Hossein Karimi [4], Mohammad Bagher Demideh[5] and Roghayeh Najjari Alamuti[6]

[1,3]Young Researchers Club, Robatkarim Branch,
Islamic Azad University, RobatKarim, Iran
abbas.akkasi@gmail.com
jafarabadm@yahoo.com
[2]Department of Computer Engineering,
Imam Khomeini International University , Qazvin, Iran
Akhaleghi@eng.ikiu.ac.ir
[4]Sama Technical and Vocational Training College,
Islamic Azad University, Yasouj Branch, Yasuj, IRAN
H.karimi.esf@gmail.com
[5]Faculty of Engineering, Department of Computer, Yasouj Branch, Islamic Azad University, Kohgiloyeh & Bovirahmad Province, Yasouj, iran
M.b.damideh@gmail.com
[6]Department of Computer Engineering,
Imam Khomeini International University , Qazvin, Iran
S936369006@edu.ikiu.ac.ir

ABSTRACT

Holding an election with aim of selecting only one person or approval / rejection of a state law, is a special kind of election which every few years in the different countries going to happen. Given the pervasiveness of this election, we must take special measures to provide high security for the referendum. Using two receipts for each voter which one is named barcode receipt, a secret indicator of vote and another is named key receipt that is a key to acknowledged the voters information box, including: voter's National Code, the candidate code which is voted by this voter, code of election station and barcode information. In this paper is proposed to enable people and social networks using data on bar code's receipts without Intrusion into the privacy of other voters, so they will put together their personal information from monitoring the election process on a social network which can help to prevent any violation in election. The security of the proposed scheme is based on the turnout in recount of votes.

KEYWORDS

Election Transparency, Electronic Referendum, Barcode Receipt

1. INTRODUCTION

There are various methods of conducting elections in the world. Some countries do their election in multi stage, like India that because of its multi billion population, has more than 10 stages for each election. [1] However, in some countries because of low population, elections are held every

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 37–41, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41204

few years. There are diversities for voting procedures, including voting at an annual meeting of the 10 persons to voting in a country that contain all people. One of the most common methods of voting is a way in which the entire population of the country with a minimum age requirement can participate in the election, and finally an option as the end result will be choose which that option has more votes in compared with other options. For example, this option can be elected a president from among 10 candidates for the Presidential Election, or approval of, or opposition to a government's decision. [1]

In this article we're going to review the election with the same Candidates for the entire country so we must select one option certainly. Accordingly, the possibility of holding such elections in each country is once or twice, annually or in every few years. For example, in Iran an election with these terms will be held once every 4 years. Therefore, any simple proposal without logical thinking and foresight, finally up to a period of several years will respond to the voting system, then because of the passage of time and the arrival of more advanced systems, its use would be without benefit plan or it will be failed. In this article we're going to have an idea for the elections in the countries that will be held in the form of a referendum, stating that at least in the next 20 years to meet the electoral needs of the country. Our favorite is the plan that makes it possible to hold the elections with the cost affordable and few times in the year, which in this situation government can get help by people's comments in critical decisions with holding an quick electronic election, It improves economic, political and social conditions of the country.[2]

2. REVIEW THE POSSIBLE ELECTION SCENARIOS

The conventional method in Iran is the same as the traditional method of election ,which we use of Fund votes and finally collect in the presence of observers, votes cast in the ballot box, are counted. the election observers (government and candidates observers) from the moment of the initial closure of the Fund's voting to stage that open ballot boxes and counting of votes ,must be present in the Polling place, given that it’s difficult to be sure that, full security for all boxes is established. In comparison to the number of infringement cases and the level of importance occurred in various elections, we have to conclude, this method is a lower security than electronic voting. [3]

Another voting methods is using of ATM machines. Given that these devices directly associated with the financial discussions, they have been designed to maximize the possibility of error to zero. In the event of an error it’s possible to reform and correct it. In the final days of the year that people are going to get the cash from these terminals, due to the increase in the number of applicants, each person must stay in the queue and wait for a long time to come his turn. If these devices used for voting, the result is that at the end of the day, a large number of voters can’t vote. This plan is good in the cases that voting days are a week or further.[4] Also by considering the importance of identification in the elections, we need at least one person as an Installing digital .observer, standing near the ATM machine, to perform the identification phase information stands in the streets is the other way of electronic voting, these devices are much lower cost than ATM machines. Because the discussion and separation of the counting of votes to count money by the banks as quite different and the voting device does not need money counter.

Another way of voting is internet voting. By considering the current equipment there isn’t solution to this issue that identification can be done via Internet and simultaneously ensuring that person would not sell his vote. Other methods of voting which will be occurred in future can vote through mobile phones, tablets and other wireless communication methods and systems. .[5]

The fundamental problem in all these cases is low possibility for running a clear election in the length of a day. Some of these cases, such as the traditional voting are possible for election in a

day, but they aren't clear enough and also the others cases have their special problems, for example an ATM is clear but the speed is not appropriate.

3. THE IDEA OF HOLDING A REFERENDUM WITH TWO RECEIPTS FOR EACH VOTE

Lack of giving receipt is a criticism that often can be seen in voting systems. Election officials refrain from offering receipt to voters that shows candidate who voted, to prevent buying and selling votes. Although this issue opposes to buy and sell votes in detail but people's confirmation of the final results of the vote counting process will disappear.

In this section, we introduce a scheme that provide receipts and ensure the transparency of the election, also the possibility of buying and selling votes in this scheme is eliminated.

It is assumed that voters use computers in the Polling place, their votes are recorded in the system. In the process of recording votes, voters logged in some basic information on the system, most important of them are the national code and the selected candidate code. Software of polling produce two receipt for each of the voter as followed below:

1. **Barcode receipt:** This is a receipt contain a barcode that have two main features. First is that based on the barcode definition each barcode receipt should be unique and second feature is that the content of this receipt, reveals that the barcode is related to a vote of the which candidate. This receipt isn't sealed.

2. **Vote key receipt:** it is a symmetric key to encode four data (voter's National code , candidate's code who selected by voter, election centre code, barcode receipt's code) The receipt is sealed by the seal of election centre.

The voter will be check the result of elections on the internet by means of barcoded receipts, obviously voter keeps the vote key receipt with himself to control in cases which he has found contradictions in votes or other existing cheats.

Anyway, the vote of voter is locked in national election centre and the key for opening it, is only in hand of voter. The locked vote (which is an encrypted text), is kept in national election centre.

3.1. The process of recording barcode receipts in the database

In the beginning of elections, election officials must set up database for candidates and register each barcode receipts of voters in the related database. All updates of these data bases are released during voting in the election centre website. For more security this database isn't online and just upload and update of it can be do online in a short intervals.

Each voters, moments after the votes can go to the website of election centre to open (or download) database about the list of his candidate votes. He must find his specific vote barcode in these databases. If the barcode does not exist in the database, it means, by anyway, his vote is ruined. In these cases, voters can inform the problem to the agent or relevant supervisor with selected candidate.
It is also possible that agents of candidates (given that they are following facts to be discovered), doing verification of voters receipts directly. It helps to speed up security check for elderly people or those who have not access to the internet

3.2 Checking the probability of selling votes election receipts

An important issue is answering to this question: whether the barcode receipt, will show that the voter has voted to whom? In the proposed scheme, two main options are designed to prevent the buying and selling of votes.

1. Barcoded receipt is printed simply without any special stamps. Presenting this receipt is just to inform that voter has vote in election, so it does not prove that it has been issued by voting system either personally(made a fake). The election verification is proved by this barcoded receipt and cheating prevention is done by vote key receipt.

2- Immediately after registration of vote, corresponding barcode releases on central website, and the person who wants to sell his barcode cannot prove to candidate's agents that this barcode in belong to his vote. Because corresponding candidate is not sure which announced number has received by himself from the voting place computer or the seller person kept the receipt of barcodes of voters which published on the website of that voting center and regarding the identity of voters are not clear , he announced himself as owner of that vote.

So with respect to above reasons it's impossible to sell a vote without confirmation stamp. Receipt of a vote is like a simple printed page and it's not validated.

3-3- possibility of producing a barcoded receipt for two person
Voter gets his receipt from machine and goes to the voting website, he sees barcode on website unaware that this receipt is possible to produce for more than one person. It means with manipulating voting software it would be possible to print repeated receipts for a candidate.

To prevent this type of problems we propose overall people supervision on process which it's free of charge and it can help safety control of election. People can send their vote's to their candidate's sms system. By discovering even a same barcode for to voters, candidate's agents can ask explanation from election organizers about occurred problem.

Election organizers should prepare facilities to vote revising. Each voter can open his own vote box by using his own key for receipt, after revising personal boxes, there are two possibilities:

1- Barcoded receipts of protests are same. In this case infraction is occurred and organizers should have a convincing answer to this problem.

2- Barcoded receipts are different. So its necessary to check whether this different receipt registered in database or not? This would show the correctness of undergoing process in voting center.

Solving the problem of producing same barcode for more than one person, would prevent consequent fraud possibilities. For example assume your receipt is not registered in database, in this case you can decide about your vote easily by using receipt of vote key.
3-4- people reports via social networks

In some countries election organizing impose so many security problems due to some economic costs and lack of supervisory facilities so that some of loser candidate claim "we were not able to detect some probable frauds because of lack of supervisors".

With growthing in social networks, it would be possible to share most part of supervisory process with people in form of their reports via sending barcoded receipts to the candiate's sms or email systems.

Proposed method prevents fraud by people help in two phase: 1) sending barcodes with any informing technique 2) collaboration with candidates in vote recounting. For example in the case of same receipts for two persons, people can help candidates with checking every changes in process of barcode production or results with presenting their own personal (with key) information.

Also there is possible to design and implement of some applications in virtual environments to virtual vote recounting or fraud detection. This type of networks can monitor vote recounting by controlling produced barcodes. Furthermore establishing some legislative rules and processes in barcode production process can help social networks for more powerful supervision.

4. CONCLUSION

Election holding in a country needs to design specific policies. If this election wants to hold just to know about global opinions about acceptance or non-acceptance of a person that would be named as "referendum". Referendum process will be differ than Senate or Municipalities election. In this paper we proposed a method to increase voting security by using 2 receipts. With keeping votes in various places and social networks controls on votes , election safety can be increased. With respect to steps of this method, the possibility of purchasing and selling votes will decrease to zero percent and all people can monitor the results at every moment. Public participation will have very important role on increasing election safety in future also it will not impose extra costs for candidates or organizers.

REFERENCES

[1] Dix A., "Electronic democracy and its implication for political privacy", 23rd International Conference of Data Protection Commissioners, September 2001, Paris,

[2] D. Balzarotti, G. Banks, M. Cova, "An Experience in Testing the Security of Real-World Electronic Voting Systems" , IEEE Transactions on Software Engineering, Vol. 36, pp. 453 - 473, May 2010

[3] S.M. Jambhulkar, J.B. Chakole, P.R.Pardhi, "A Secure Approach for Web Based Internet Voting System Using Multiple Encryption", International Conference on Signal Processing and Computing Technologies(ICESC), vol 4, pp. 371-375, jan 2014

[4] Jefferson, D. and Rubin, A. and Simons, B. and Wagner, D. A Security Analysis of the Secure Electronic Registration and Voting Experiment (SERVE), Online, Available from http://www.servesecurityreport.org/, last accessed 2014.

[5] Aviel D. Rubin. Security Considerations for Remore Electronic Voting. Communications of the ACM, 45(12), 2012.

INTENTIONAL BLANK

MEASURING SIMILARITY BETWEEN MOBILITY MODELS AND REAL WORLD MOTION TRAJECTORIES

Morteza Mousavi Barroudi

Student Member, IEEE, Aaron Harwood, Shanika Karunasekera

ABSTRACT

Various mobility models have been proposed to represent the motion behaviour of mobile nodes in the real world. Selection of the most similar mobility model to a given real world environment is a challenging issue which has a significant impact on the quality of performance evaluation of different network protocols. In this paper we propose a methodology for measurement of similarity between mobility models used in mobile networks simulation and real world mobility scenarios with different transportation modes. We explain our mobility metrics we have used for analysis of motion behavior of mobile nodes and a pre-processing method which makes our trajectories suitable for extraction and calculation of these metrics considering shape of the road networks and GPS noise. Then we use a feature selection method to find the most discriminative features which are able to distinguish between trajectories with different transportation modes using a supervised learning and feature ranking method. Subsequently, using our selected feature space we perform Fuzzy C-means Clustering to find the degree of similarity between each of our mobility models and real world trajectories with different transportation modes. Our methodology can be used to select the most similar mobility model suitable for simulation of mobile network protocols (such as DTN and MANETs protocols) in a particular real world area.

KEYWORDS

Mobility Models, Similarity Analysis, Transportation Modes, Mobile Networks

1. INTRODUCTION

Researchers often use simulators such as Glomosim, Opnet, NS2, and OMNET++ to simulate their mobile networks protocols. These simulators provide facilities for simulation of motion of mobile nodes in the plain simulation area using mobility models such as Randowm Waypoint, Random Walk, Brownian Motion [1], Markovian [2], and RPGM [3], [4], [2]. One question here is which of these highly used mobility models perform more similar to real world motion scenarios? To be able to find the best mobility models for simulation and performance evaluation of mobile network protocols we need to perform similarity analysis between mobility models and real world motion behaviour of mobile nodes in the particular real world area in which we are going to implement our networking protocol.

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 43–55, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41205

Research works such as [3], [4] and [5] proposed mobility models for simulation of mobile node motion in mobile

Morteza Mousavi Barroudi is with NICTA VRL, Department of CIS, The University of Melbourne, Melbourne, Australia. email: mousavi@student.unimelb.edu.au.

Aaron Harwood is with NICTA VRL, Department of CIS, The University of Melbourne, Melbourne, Australia. email: aharwood@unimelb.edu.au.

Shanika Karunasekera is with NICTA VRL, Department of CIS, The University of Melbourne, Melbourne, Australia. email: karus@unimelb.edu.au.

National ICT Australia (NICTA) is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council. environments. However, they did not provide a comprehensive methodology for analysis of the similarity of their proposed mobility models and real world scenarios with different transportation modes, ("Bike", "Car", "Train", "Walk", "Bus").

In this paper we extend the methodology we introduced in our previous work [6] to propose a comprehensive methodology for measuring the similarity between real world movement scenarios with different transportation modes and mobility models. Aiming that we perform a feature selection method to find the most discriminative feature sets which are able to classify different trajectories with different transportation modes into their right classes with optimal set cardinality [6]. Subsequently, we propose a similarity measurement method which is able to calculate the degree of similarity between mobility models and real world trajectories with different transportation modes.

Before performing the similarity analysis method, we need to pre-process our data to make it suitable for feature extraction. We use our proposed method in [7] to interpolate missing waypoints (up-sampling) and reduce GPS noise to have regular and reliable real world trajectories. Finally, to calculate the degree of similarity between mobility models and real world transportation modes, we use Fuzzy C-means Clustering (FCM) [8]. This clustering method calculates the degree of membership of each sample in the feature space to each of the clusters. We use this fuzzy membership as a measure for the similarity between our samples (trajectories with different transportation modes) and clusters of mobility models. Before, performing the clustering we extract features from 5 classes of mobility models, (Random Waypoint (RWP), Random Walk (RW), Levy Walk(LW), Manhattan (MAN), and RPGM [6], [3], [4], [2], [5]) as representatives of mobility models. We also extract the features of our transportation mode trajectories. Then we use the optimal and the most discriminative feature set calculated in the feature selection phase as the feature space in our FCM clustering and consider the estimated degrees of membership as the degree of similarity between each of the transportation modes and each of the mobility models.

The remaining of the paper is organized as follows. In section II, we shortly discuss the most related research works to ours. In Section III, we briefly introduce our mobility metrics. Then we introduce our proposed trajectory pre-processing, classification and feature selection methods in IV-B. IV-D introduces our similarity analysis method using fuzzy Cmeans clustering. Finally, Section V contains the conclusion and future work.

Table I. Mobility Metrics

1	Degree of Direction AutoCorrelation
2	Degree of Speed AutoCorrelation
3	Entropy of Direction Probability Distribution
4	Entropy of Direction Change Probability Distribution
5	Entropy of Position Density Probability
6	Average speed
7	Average Acceleration
8	Speed Change Rate
9	Direction Change Rate
10	Stop Rate
11	Speed Variance

2. RELATED WORKS

The most related works to ours are works done on analysis the nature of human mobility such as [9], [5], [10], [11]. To the best of our knowledge the most comprehensive work in this area is [5] in which researchers have analysed probability distribution of flight length, pause time, flight speed, and mean squared displacement of human mobility. They have used the data collected from 5 different places and done statistical analysis to find the best probability distribution for each of the above features. Then they have used the results of the statistical analysis to propose a mobility model called Levywalk.

Although [5] is one of the best works done on analysis on human mobility, there are some drawbacks in the work (which we are aiming to deal with in this paper) as follows, they have used linear interpolation for estimating the missing GPS waypoint and have not considered the GPS noise and the shape of the road networks [12] and their impact on the quality of their statistical analysis. Moreover, they have not considered different transportation modes in their analysis. Furthermore, they have not analysed the discriminative power of each of the metrics they have proposed to distinguish between real world trajectories. In addition, they have not proposed a comprehensive methodology for pattern recognition and analysis of the similarity between real world trajectories and mobility models using different mobility metrics.

3. MOBILITY METRICS

In order to analyze mobility trajectories, we need to extract features that represent their actual motion behavior in real world environments. For extraction of these features we need to define some mobility metrics of motion trajectories. In this section we briefly introduce our mobility metrics we have used in our feature extraction method [6], [7], [13].

Table I assigns a number to each of the mobility metrics to achieve simplicity in naming them.

A. Degree of Direction Autocorrelation

This metric examines the degree of temporal dependence between a node's direction at the current sample time and Δt sample time earlier [6]. We take this time difference to be 1 sample time [6].

$$\mathrm{DAC}(i,t,t') = \mathrm{RD}(\vec{v_i}(t), \vec{v_i}(t')) = \cos(\theta),$$

where RD is the cosine of angle between two vectors given by:

$$\frac{\vec{a}(t).\vec{b}(t)}{\|\vec{a}(t)\|\|\vec{b}(t)\|}$$

The Average Degree of Direction Correlation, which is the average of Direction Correlation over all nodes and all time instances in a mobility trace, is calculated as follows [6]:

$$\overline{\mathrm{DAC}} = \frac{\sum_{t=1}^{T} \mathrm{DAC}(i,t,t')}{T}.$$

B. Degree of Speed Autocorrelation

This metric examines the degree of temporal dependence between a node's speed from the current sample time to Δt sample time earlier [6]. We take this time difference to be 1 sample time [6].

$$\mathrm{SAC}(i,t,t') = \mathrm{SR}(\vec{s_i}(t), \vec{s_i}(t')),$$

where SR is the speed ratio between two vectors, is given by

$$\mathrm{SR} = \frac{\min|\vec{s_i}(t), \vec{s_i}(t')|}{\max|\vec{s_i}(t), \vec{s_i}(t')|},$$

The Average Degree of Speed Auto Correlation, which is the average of Degree of Speed Auto Correlation over all time instances in a trajectory, is calculated as follows [6]:

$$\overline{\mathrm{SAC}} = \frac{\sum_{t=1}^{T} \mathrm{SAC}(i,t,t')}{T},$$

where *P* is the number of tuples (*n, t, t`*).

C. Average Speed

The speed of a mobile node is obviously one useful feature for separation of transportation modes [13]. To extract this feature we have calculated the average of speeds which are more than a certain threshold (0.2 m/s). We did this to eliminate the impact of zero speeds in the average speed. This feature has high value in "Car", "Train" and "Bus" and lower value in "Bike" and "Walk" [13].

D. Speed Variance

Speed variance shows the variance of speed of a mobile node during its trajectory time period [13]. This feature has high value in "Car", "Train" and "Bus" and lower value in "Bike" and "Walk" as well.

E. Average Acceleration

Acceleration variance shows the variance of speed of a mobile node during its trajectory time period. Since the acceleration can have positive and negative values we have used the absolute value of acceleration to have a better view of acceleration of our trajectories [13].

F. Direction Change Rate

Direction or heading change rate indicates how frequently a mobile node changes its heading direction during their trajectories [13]. This feature is higher in "Walk" because people are not usually as restricted to road networks (with constant direction in a straight road segment) as cars and buses are.

This value is rather higher in "Bike" as well. To calculate this feature we have found the number of the heading direction changes (DC) which are larger than a certain threshold. Then we normalize this value, for each trajectory, by dividing it by distance or length of the trajectory (in meters).

DCR = DC/*TrajectoryDistance*:

G. Speed Change Rate

In trajectories with transportation modes such as "Walk", "Bike" and "Bus" the the frequency of speed changes are higher than others [13]. To calculate this feature, we have considered speed changes which are more than a certain threshold as a speed change. Then we have divided the number of speed changes during a trajectory time period by length of the trajectory similar to above.

H. Stop Rate

In trajectories with transportation modes such as "Walk", "Bike" and "Bus" the frequency of stops are higher than others [13]. Therefore, this feature is a good feature to distinguish between different transportation modes. To calculate this feature, we have found the frequency of stops and divided it by the length of trajectories similar to above.

I. Entropy of Position Probability Distribution

Position Density measures how mobile nodes are distributed in the analysis area of each trajectory. To generate this metric we first find the location density distribution for each of the trajectories and then find the entropy of the probability distribution. This metric would be high in trajectories where mobile nodes are located in specific areas of a map, considerably more than at other areas, i.e. in nonuniform distributions, and low in traces where nodes are uniformly

distributed [6]. For example, in the traces with bus transportation mode, since mobile nodes are restricted to streets in an urban map, most of the time their position distribution density is much higher in street areas than building blocks (compared to walk transportation mode) and similarly higher in intersections and bus stops than streets. Therefore, this metric would be appropriate to distinguish trajectories that have more geographical restrictions [6]. Note that, in our previous work, we have used variance of position densities instead of entropy which is less discriminative.

To generate the metric, first we extract position density of mobile nodes in each trajectory throughout the trajectory period, by transforming our trajectory coordination system from geodetic to Cartesian, dividing our trajectory areas into cellular areas and calculating the density of the mobile nodes in each cell inside the grid environment of each trajectory.

The Position probability in each of the cells is computed as follows:

$$P(Exists(i,j)) = \frac{\sum_{t=1}^{T} Exists(t,i,j)}{T},$$

where Exists(*i, j, t*) = 1 if node n exists in Cell (*i, j*) at time *t*, and otherwise 0.

In other words, position probability is defined as the probability of existence of a mobile nodes in a particular cell in a specific time. Then we calculate Entropy of Position Probability Distribution as follows:

$$H(\text{Exists}) = -\sum_{\text{I}=1}^{I}\sum_{\text{J}=1}^{J} P(\text{Density(i,j)}) \log_2(P(\text{Density(i,j)})),$$

where *i* and *j* are coordinates of cells in the simulation area. In our analysis *I* and *J* are set according to width and height of our trajectory environment

J. Entropy of Direction Change Probability Distribution

This metric allows us to compare different types of transportation modes. It shows the degree of uniformity in the direction changes of a mobile node during its trajectory period. In an urban area, vehicles and people are restricted to streets and road networks. The turning angles of vehicles are restricted mostly to some particular angles. For instance, in urban areas, objects are not able to turn and change their directions with a random angle. They are restricted to turn based on the road network conditions. In contrast, an animal in a farm is often able to move almost in any direction or turn with any angle. Therefore, density of direction changes is substantially non-uniform in urban areas in contrast areas with no geographical restrictions.

To extract direction change probability distribution of a trajectory, we use our method propose in [7]. Firstly, we calculated direction changes of the moving object in each of the trajectories. Then we calculated the probability of each of the direction change angles between 1 and 180 degrees. Thus, we generated a list of 180 direction change angles and their probabilities. We have rounded the direction changes to have a discrete set of direction changes {1; 2; :::; 180}. Probability for

direction change with angle β is defined as $\acute{P}(DC_\beta) = \frac{DC_\beta}{A}$. DC_β is the count of direction changes with angle β and A is the count of all direction changes in the trajectory [7].

We propose Entropy of Direction Change probability Distribution as a numeric measure to compare direction change probability distributions of trajectories with different transportation modes [7]. In information theory and probability, entropy [14] is a measure of amount of information in a signal. It indicates amount of disorder and degree of uncertainty of a probability distribution as well. Shannon's Entropy of direction change probability distribution of a trajectory is defined as follows [7]:

$$H(\mathrm{DC}) = -\sum_{\alpha=1}^{180} P(\mathrm{DC}_\alpha)\log_2(P(\mathrm{DC}_\alpha)).$$

K. Entropy of Direction Probability Distribution

In urban environments in accordance with the shape of the map of road networks, vehicles and people usually move with some particular directions much more than others. As a result the probability distribution of direction angles of moving objects is not uniform. Degree of uniformity or degree of disorder of direction probability distribution can be considered as another measure for comparison of trajectories with different transportation modes. To be able to analyze the behavior of our trajectories, we calculated the entropy of direction probability distribution of trajectories, similar to entropy of direction change probability distribution [7] discussed in III-J with the difference that here we have 360 rounded discrete direction angles {1; 2; :::; 360}. We define Shannon's Entropy of direction change probability distribution of a trajectory as follows [7]:

$$H(\mathrm{DA}) = -\sum_{\alpha=1}^{360} P(\mathrm{DA}_\alpha)\log_2(P(\mathrm{DA}_\alpha)).$$

4. SIMILARITY ANALYSIS

In this section we briefly introduce the steps of our similarity analysis methodology. Our purpose is to find the degree of similarity between each of the real world transportation modes to each of our mobility models. Aiming for that, we first pre-process our real world trajectories to make them suitable for feature extraction (IV-A). Secondly, we find the best feature sets which are able to discriminate and distinguish different transportation modes from each other (IV-B) using a supervised learning method. Subsequently, using the best feature sets we perform the Fuzzy C-means clustering method to find the degree of membership of each of the transportation modes to each of our mobility model clusters and consider this membership as degree of similarity between mobility models and transportation modes (IV-D).

As mentioned before, selection of the similarity metrics is a major issue in any similarity analysis because the degree of similarity is highly dependent on the feature space we are using in our machine learning process [15]. We use the best feature set selected in the feature selection phase discussed in IV-B as the feature space in the FCM method because we have found that it is the

optimum feature set that can distinguish trajectories with different transportation modes from each other with maximum accuracy.

A. Trajectory Pre-processing

We have used a very large database of trajectories collected by Microsoft Research in Beijing, China [16], [17], This dataset contains trajectories of 181 people moving for almost 3 years. We inserted every trajectory with known transportation mode into our database. Then, we selected 100 trajectories (with highest GPS sampling rate) of each of the transportation modes. Hence we have a dataset of trajectories of different mobile nodes with known transportation modes.

To be able to extract the metrics introduced above, we needed to process our raw GPS trajectories to make them suitable for feature extraction. Due to irregularity and low ampling rate in trajectories collected by GPS, we do not have access to data about position, speed and direction of our mobile nodes at all needed time samples. For example, in our processing we need to compute the direction autocorrelation of each mobile node once every 10 seconds. However, the GPS trajectories sampling rate is not regular and do not include spatio-temporal information of the mobile nodes regularly for all needed times. Moreover, particularly in urban environments we need to provide map-matching [12], [18], [19] to reduce GPS noise and make our trajectories more dependable. In previous related works such as [5], linear interpolation has been applied on the raw GPS trajectories to interpolate missing waypoints (at needed sampling times). However, linear interpolation extremely suffers from inaccuracy [7].

We have used our previously proposed method called map based spatio-temporal interpolation [7] to interpolate missing waypoints and do map-matching on our trajectories. Map based interpolation uses real world maps and the estimated speed for each road segment to estimate the position of each mobile node at each given query time on the road network. We have called HSTQ query [7] for each needed sampling time and up-sampled our real world trajectories to archive regularly sampled GPS trajectories with reduced GPS noise. As opposed to linear interpolation used in [5], map-based interpolation has very higher accuracy in estimated positions and estimated turning angles and speeds. This method implicitly does mapmatching [12], [20] to reduce GPS noise as well.

After performing the pre-processing phase, we have 100 trajectories for each of our transportation modes (“Walk”, “Car”, “Bike”, “Train”, “Bus”) extracted from our dataset [16] with regular sampling rate and reduced GPS noise suitable enough for our feature extraction phase. Then we separated the trajectories into two sets of training and test trajectories for each of the transportation modes.

B. Classification and Feature Selection

The goal of the feature selection [15] process is to find the best (most discriminative) feature subsets (comprised of calculated values of mobility metrics) which are able to distinguish between each specific transportation mode from other transportation modes. The feature set should be optimal. In other words, we should find feature sets with the highest accuracy and minimum set cardinality. For example if we can use only speed Autocorrelation mobility metric to distinguish between “Walk” and “Car” transportation modes, it does not make sense to run the feature extraction and classification using all the mobility metrics. Therefore, we choose the most

discriminative feature set as our similarity analysis feature set among so many different possible feature sets (in our study it would be $2^{11} = 2048$ different feature sets) [6].

Depending on our trajectories and our environment conditions (shape of the road network and traffic conditions, etc), the best feature set selected by our feature selection method may slightly change. Therefore, we should perform all the steps proposed in this paper to find the best feature set and find the degree of similarity between our mobility models and real world trajectories in that particular environment.

Table II. Selected Optimal Feature Sets

Transportation Mode	Feature Sets	Accuracy (%)
All	{1,3,4,9}	74.4%
Walk	{7,8,9}	74%
Car	{11}	98%
Bus	{1,6,11}	48%
Bike	{1,3,4,8,10}	78%
Train	{7,11}	98%

1) Feature Extraction: We performed the feature extraction phase resulting in two training and test tables each of which include 250 rows and 11 columns (mobility metrics in tableI). We also generated test tables for each of the transportation modes separately (each including 50 rows) to be able to analyse performance of our feature sets for each of the transportation modes separately.

2) Classification: We used the K-nearest neighbor (KNN) [15] classification method to classify each of our samples in the test table to it's nearest classes. We trained the classifier with our labeled training samples. Then we classify the test samples into their nearest classes. Euclidian distance was used for finding *k* nearest neighbors and parameter K was set to 10 [6].

3) Feature Ranking: To find the optimal feature set, firstly, we generated all possible feature subsets of our feature set. Subsequently, we tested the accuracy of each of the feature subsets in classification of different transportation modes. Then we ranked each feature set using its classification accuracy combined with cardinality of feature set [6].

4) Results: Table II shows selected most discriminative and best feature sets for classification of each of transportation mode. The word "All", means we have used every 250 test trajectories for classification. As it is clearly seen, the best feature set which is able to distinguish all the transportation modes from each other with highest accuracy (74:4%) and lowest set cardinality (4) is {1; 3; 4; 9}. Suppose we have a trajectory with unknown transportation mode, our experimental results suggest that if we need to classify it into its right transportation mode class we should use features {1; 3; 4; 9} (feature space comprised of degree of direction autocorrelation, entropy of direction probability distribution, entropy of direction change probability distribution, and direction change rate) and make a 4 dimensional feature space for our classification.

Based on results depicted in table II, for example, if we need to classify trajectories with "Walk" transportation mode into right class ("Walk" class) we should use feature set comprised of speed change rate and direction change rate features.

By using this feature set if a trajectory had a "Walk" transportation mode, it will be detected with 98% accuracy and there is 2% chance for it being classified wrongfully. We can infer the same results from other rows in table II similarly.

C. Mobility Models Simulation Configurations

To be able to preform similarity analysis using the selected feature set, ({1; 3; 4; 9}), we generated one mobility trace for each of the following mobility models that we needed to analyse: (RW, RWP, RPGM, MAN, and LW) using our previously developed mobility simulator [2]. Each of the mobility traces include 50 mobile nodes moving in an area of 10K X 10K meters moving for 10000 sample times. Minimum and maximum speed in all the models are set as 1 m/s and 25 m/s respectively. Maximum pause times for RWP and Manhattan have set to 10 sample times and minimum pause time to 1 sample time. We used a regular Manhattan map for the Manhattan mobility model with 50 equidistant intersections in the simulation area [2]. For Levy Walk mobility model, we set the maximum pause time to 10 and minimum pause time to 1 sample times. We set arameters α to 1, max flight to 100 meter, min flight 30 meter and parameter β to 1 [5]. For RPGM mobility model, the central node moves with RW model. The *sdr* and *adr* parameters in RPGM have been set to 0:05. Walk time parameter in RW has been set to 20 sample times.

Then we performed the feature extraction similar to the way discussed in IV-B1.

D. Fuzzy C-means Clustering

To perform similarity analysis we use Fuzzy C-means clustering algorithm [8]. FCM is a clustering method which in addition to estimating the cluster to which a sample belongs, reports the degree of membership of a sample to all the clusters. We use the reported memberships ([0; 1]) converted to percentage as a measure of similarity of a sample to a cluster.

We generated a table comprising of 250 rows and 11 column (metrics). We also used the same (real world) trajectories used as the training set in IV-B. For simplicity, we used the average of features extracted form each of the group of the transportation modes as the representative of each of the transportation modes. So, for each of the transportation modes we had a new table with 251 rows. The last row contains average value for each metric for the particular transportation mode. We used the FCM method using MATLAB on our extracted data. The number of desired clusters in FCM has been set to 5 (the count of the transportation modes).

Table III shows the results provided by FCM clustering method with the configurations discussed above. It shows the degree of similarity between each of the transportation modes and each of our considered mobility models (in percent).

Although the purpose of this paper is to propose a comprehensive similarity analysis ethodology, not to find the best mobility models for simulation, one interesting outcome of our experimental results shows that Levy Walk [5] is the best mobility model among our mobility models in four of the transportation modes. This similar results to [5] can confirm the performance of our similarity measurement method and also confirm the performance of LW mobility model to mimic the real world mobility using new mobility models and more comprehensive analysis. On the other hand, the Random Waypoint mobility model, although being one of the most highly used mobility

models used in mobile networks simulations such as DTN and MANET protocols [3], [4], has much lower performance in comparison with Levy Walk; based on our experiments using our selected features. Random walk which is very similar to Brownian motion also does not perform well in most cases.

Table III. Degree of Similarity

Transportation Mode	Mobility Model	Degree of Similarity (%)
Car	RW	54.25%
	LW	36.84%
	RWP	4.35%
	RPGM	2.62%
	MAN	1.75%
Bike	LW	43.34%
	RW	29.69%
	RWP	11.80%
	RPGM	9.65%
	MAN	5.52%
Bus	LW	56.66%
	RWP	22.51%
	RW	11.81%
	MAN	5.09%
	RPGM	3.93%
Train	LW	62.26%
	RW	27.72%
	RWP	5.60%
	RPGM	2.44%
	MAN	1.98%
Walk	LW	33.87%
	RW	23.88%
	RWP	18.51%
	RPGM	13.31%
	MAN	10.43%

5. CONCLUSION AND FUTURE WORK

In this paper we proposed a comprehensive methodology for measurement of similarity between mobility models often used in simulation of mobile networks protocols and real world motion trajectories with different transportation modes. We introduced metrics and a method for classification of trajectories with different transportation modes. Subsequently, we briefly introduced a pre-processing method we need to perform to achieve GPS trajectories with regular sampling rate and reduced noise. Then we used a feature ranking method and found best features for classification of each of the transportation modes from others. Then we provided a method for similarity analysis using Fuzzy C-means clustering method. Depending on our trajectories and environment of our networking protocol, the best feature set selected by our feature selection method may change. Hence, we need to perform all the steps proposed in this paper to find the best feature set and find the degree of similarity between our mobility models (which can be

different to the ones we have considered) and real world trajectories (with other kinds of transportation modes in different area with other road networks and traffic conditions).

As future work, we will consider new features such as frequent and periodic behaviour to distinguish between similar transportation modes such as cars and buses using time series similarity analysis. Moreover, we will work on proposing mobility models which are able to simulate frequent and periodic behaviours other than spatio-temporal, geographical behaviours.

REFERENCES

[1] S. Ioannidis and P. Marbach, "A brownian motion model for last encounter routing." in INFOCOM. IEEE. [Online]. Available: http://dblp.uni-trier.de/db/conf/infocom/infocom2006.htm

[2] S. M. Mousavi, M. Moshref, H. R. Rabiee, and A. Dabirmoghaddam, "Mobisim: a framework for simulation of mobility models in mobile ad-hoc networks," in Proceedings of 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications. New York, USA, October 2007.

[3] F. Bai, N. Sadagopanb, and A. Helmy, "The important framework for analyzing the impact of mobility on performance of routing protocols for adhoc networks," Ad Hoc Networks, vol. 1, no. 4, pp. 383–403, November 2003.

[4] T. Camp, J. Boleng, and V. Davis, "A survey of mobility models for ad hoc network research," Wireless Communication and Mobile Computing, vol. 2, no. 5, pp. 483–502, August 2002.

[5] I. Rhee, M . Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility," Networking, IEEE/ACM Transactions on, vol. 19, no. 3, pp. 630–643, June 2011.

[6] M. M. Barroudi, A. Harwood, and S. Karunasekera, "Feature selection for user motion pattern recognition in mobile networks," in PIMRC, 2012, pp. 1521–1527.

[7] ——, "Map-based spatio-temporal interpolation in vehicle trajectory data using routing web-services," in Proceedings of the 5th ACM SIGSPATIAL International Workshop on Computational Transportation Science, ser. IWCTS '12. New York, NY, USA: ACM, 2012, pp. 43–48. [Online]. Available: http://doi.acm.org/10.1145/2442942.2442951

[8] J. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, no. 2-3, pp. 191–203. [Online]. Available: http://dx.doi.org/10.1016/0098-3004(84)90020-7

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," Nature, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[10] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: Self-similar least-action human walk," IEEE/ACM Trans. Netw., vol. 20, no. 2, pp. 515–529, Apr. 2012. [Online]. Available: http://dx.doi.org/10.1109/TNET.2011.2172984

[11] D. Brockmann, L. Hufnagel, and T. Geisel. (2006, May) The scaling laws of human travel. [Online]. Available: http://arxiv.org/abs/condmat/0605511

[12] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current mapmatchingalgorithms for transport applications: State-of-the art and future research directions," Transportation Research Part C: Emerging Technologies, vol. 15, no. 5, pp. 312 – 328, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X07000265

[13] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in Proceedings of the 10th international conference on Ubiquitous computing, ser. UbiComp '08. New York, NY, USA: ACM, 2008, pp. 312–321. [Online]. Available: http://doi.acm.org/10.1145/1409635.1409677

[14] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 4th ed., 2002.

[15] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. Wiley,2001.

[16] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," IEEE Data Eng.Bull., vol. 33, no. 2, pp. 32–39, 2010.

[17] —, "Mining interesting locations and travel sequences from gps trajectories," in WWW 2009. Association for Computing Machinery, Inc., April 2009, wWW 2009. [Online]. Available:http://research.microsoft.com/apps/pubs/default.aspx?id=79440

[18] Y. Zheng and X. Zhou, Eds., Computing with Spatial Trajectories.Springer, 2011..

[19] G. Trajcevski, "Uncertainty in spatial trajectories," in Computing with Spatial Trajectories, 2011, pp. 63–107.

[20] J. Krumm, "Trajectory analysis for driving," in Computing with Spatial Trajectories, 2011, pp. 213–241.

INTENTIONAL BLANK

SEPS-AKA: A SECURE EVOLVED PACKET SYSTEM AUTHENTICATION AND KEY AGREEMENT SCHEME FOR LTE-A NETWORKS

Zaher Jabr Haddad[1], Sanaa Taha[2] and Imane Aly Saroit Ismail[2]

[1]Department of Computer Science, Faculty of Applied Science,
Al-Aqsa University, Gaza, Palestine
zj.haddad@alaqsa.edu.ps
[2]Information Technology Department, Faculty of Computers and Information,
Cairo University, Cairo, Egypt
staha@fci-cu.edu.eg, i.saroit@ fci-cu.edu.eg

ABSTRACT

In this paper, we propose a secure authentication of the Evolved Packet System Authentication and Key Agreement (EPS-AKA) for the LTE-A network. Our scheme is proposed to solve the problem of sending the IMSI as a clear text, and hence prevents the mobility management entity attack. We will use public key (PK) cryptography to protect the transmitted messages, the RSA scheme computation to compute a temporary value to the IMSI, and nonce to generate challenge messages toward the opposite side. Our scheme does not need to change the original framework and the infrastructure of the LTE-A network, although a ciphered IMSI is transmitted. The authentication procedure is performed by the HSS to authenticate the UEs and the MME; therefore, the impersonating of the MME and UEs is not possible. Our evaluation demonstrates that the proposed scheme is secure and achieves the security requirements of the LTE-A subscribers such as privacy, authentication, confidentiality and integrity. In our scheme, we try to maintain the problems defined in the previous related works.

KEYWORDS

Long Term Evolution – Advanced, Authentication and Key Agreement, Home Subscriber Server, Mobility Management Entity, User Equipment.

1. INTRODUCTION

The Long Term Evolution-Advanced (LTE-A) network is a packet based system specified by the Third Generation Partnership Project (3GPP) towards fourth-generation (4G) mobile; in order to meet more subscriber needs. Among those communications, LTE-A is the next generation of the cellular communication system that meets more subscriber needs, such as: 1) wider bandwidth that supports up to 100MHz via aggregation of 20 MHz blocks, 2)Multi Input Multi output (MIMO) that allows the use of multiple antennas in the transmitter and the receiver in order to

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 57–70, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41206

improve the communication performance, 3) coordinate Multiple transmission (CoMP) that allows coordinates scheduling, beam-forming and joint processing transmission, 4)Heterogonous Network (Het-Net) that supports enhanced inter-cell interference coordinate (eICIC) to deal at the interference issues at the cell edge, and 5)relaying capabilities that achieves self-backhauling of the radio signal between a base station and User Equipment (UE) [1].

The EPS-AKA is the authentication protocol used by the LTE-A network to perform the authentication and key agreement security services. EPS-AKA protocol was improved to prevent malicious attacks such a redirection, rogue base station, and Man in the middle attacks. A malicious attack can be any action intents of acquiring, destroying, modifying or accessing a transmitted data without permission. However, the lack of privacy and denial of services attack still a big weakness of the EPS-AKA protocol. This LTE-A's security weakness is represented in the processes of registration, synchronization failure, and roaming to a new mobility management entity (MME), when the MME requests the international mobile subscriber identity (IMSI) of the User Equipment (UE). Therefore the IMSI disclosure may incur severe problems [2].

Many attacks may violate the vulnerabilities of the authentication in the LTE-A network such as [3][4]:

a) Replay Attack, which attains to perform maliciously or fraudulently repeated or delayed transmitted messages in order to increase the flow in the network and therefore, may make system toppled [3][4].
b) Denial of service (DoS) attack, which attains to make a machine or network resources unavailable to legitimate users [3][4].
c) Man in the Middle (MITM) attack, which makes independent connection between two victims in order to intercept or inject fake messages [3][4].
d) Impersonation attack, which attempts to use a fake identity to gain unauthorized access to network system through legitimate access identification [3][4].

In this paper, a novel scheme is proposed to solve the problem of sending the IMSI as a clear text, and hence prevents the mobility management entity attack. In our scheme, we will use three levels of security. First, nonce is used to generate challenge messages toward the opposite side. Second, PK cryptography is used to protect the transmitted messages. Third, the RSA scheme computation is used to compute a temporary value to the IMSI.

The remainder of this paper is organized as follows: Section II illustrates the related work. The system models, including network, threat, and trust models, are presented in Section 3. In section 4 the preliminaries are discussed. In section 5, the proposed system, SEPS-AKA, is introduced. In sections 5 and 7, the security analysis and performance evaluation are discussed, respectively. In section 8, conclusion and future work are provided.

2. RELATED WORK

In [5], Park et al., introduce number of possible security risks may be caused due to the open nature of the 4G networks. First, a large number of external connectivity points with peer operator, third-party applications providers, the public Internet, and with numerous heterogeneous technologies accessing the infrastructure, serves as potential security holes if the security technologies do not fully interoperability. Second, multiple service providers share the core network infrastructure, meaning that compromise of a single provider may result in collapse of the entire network infrastructure. Third, service theft and billing fraud can take place if there are third-parties masquerading as legitimate ones [5]. New end-user equipment's can also become a source of malicious (e.g., DoS) attacks, viruses, worms, spam mails and calls. In particular, the

Spam over Internet Telephony (SPIT), the new spam for VoIP [5], becomes a serious problem just like the e-mail spam today. For example, SPITs targeting VoIP gateways can consume available bandwidth, thereby severely degrading QoS and voice quality. Clearly, the open nature of VoIP makes it easy for the attackers to broadcast SPITs similar to the case of spam emails. Other possible VoIP threats include: (1) spoofing that misdirects communications, modifies data, or even transfers cash from a stolen credit card number, (2) SIP registration hijacking that substitutes the IP address of packet header with attacker's own, (3) eavesdropping of private conversation that intercepts and crypt-analyzes IP packets, and (4) phishing attacks that steal user names, passwords, bank accounts, credit cards, and even social security numbers.

In [6], Purkhiabani et al. propose a new scheme to preserve the privacy of the IMSI by encrypting it using a temporary random number (MSR) during its transmission into the EUTRAN interface. In addition, HSS generates only one Authentication Vector (AV) to use in each authentication process in order to preserve the bandwidth of the CN. Therefore, the UE sends a message, which contains a concatenation of MSR, IMSI, and MSMAC, to the MME, which in turn forwards this message to the HSS. The MSR is a random number generated by the UE, and MSMAC= $f1_k$(MSR), where f1 is a cryptographic function used to generate 128-bit output using 128-bit input key. After receiving the transmitted message, the HSS verifies the IMSI, generates and sends back one AV to the MME, and hence, the original authentication and key agreement are performed. In this scheme, the use of same framework of the original LTE-A authentication scheme, decreases the HSS bandwidth consumption, the protection of the IMSI, and bandwidth preserving of the CN. But as the MSR is generated by the UE, this increases the possibility of malicious UE and MME. Moreover, the bandwidth consumption is moved from the CN to the radio interface.

In [7], Hamandi et al. propose a scheme to solve the privacy problem in the LTE-A authentication scheme to prevent the masquerading of the MME. Authors employ the public key infrastructure to provide more powerful MME and HSS elements. MME generates and sends a random number, RANDMME, to the UE to compute a vector of parameters, and then returns a message to the MME, which in turn adds its identity and digital signature and forwards the message to the HSS. At receiving, the HSS verifies the identities of both the MME and IMSI, and then generates a new random mobile subscriber identity, RMSI, to concatenate with the AVs. Both AVs and RMSI are sent back to the MME in order to complete the original authentication and key agreement procedures. In this scheme, a ciphered IMSI is sent, a virtual number (TMSI) is used in the next hops instead of the IMSI, and legal MME identity is protected by digital signature. However, EUTRAN consumption is increased. In addition, the initiation procedure is started from the MME, which also allows the possibility of the presence of malicious MME. Additionally, this scheme is not integrated with the original mobility procedures, such as handover and localization [7].

In [8], Abdo et al. define four security weaknesses in the original LTE AKA protocol: IMSI catching, tracking user temporary identity due to linkability and security network authentication. In addition, the authors propose two countermeasures to use in order to solve these problems: Public Key Infrastructure (PKI) and pseudonyms based approaches. The advantage of this work is the security capabilities that are performed using the PKI. However, there is a critical problem that is the first hop dependency, where the UE depends on a pre-stored cipher Key (CK) and identity Key (IK) to generate the initial pseudonyms. CK and IK are generated by the pre-shared cryptographic function using the pre-shared secret key (K) between UE and HSS and a random challenge RAND that is generated by the HSS, therefore, the HSS should performs some computations before initialization phase, and surely this depends on the IMSI.

In [9], Abdo et al. propose a scheme called EPS mutual authentication and Crypt-analyzing (SP-AKA), which is a self-certified based protocol, in order to solve the positive capturing of the IMSI during user identification and key agreement protocols. Authors use the PKI to encrypt the transmitted AKA messages. Hence, provide a high security level, but the fake MME is still a problem.

In [10], Lai et al. propose a new scheme for group base communication authentication called a secure and efficient group authentication and key agreement protocol for LTE networks (SE-AKA). This scheme uses the Elliptic curve Diffie-Hellman to achieve the key forward/backward secrecy and also adapts asymmetric cryptosystem to protect user privacy. For group authentication, SA-AKA uses a group temporary key (GTK), which employs a well-known keys generation algorithm, Diffie-Hellman, and also provides a strong security level where subscriber must meet the restrictions of the authenticated group, before network authentication. The problems of this scheme are the consumption of the MME where the Elliptic curve Diffie-Hellman consumes time to generate and distribute the public keys between group members, while the main role of the MME is to work as a gateway between the HSS and the UEs. Also, the proposed group is considered as an uncontrolled area in the network and used to break the security of the network since the authentication permissions are invoked to the group authority instead of the HSS.

In [11], Zheng el at. propose a hybrid AKA scheme that uses a trusted model platform and PKC to adapt the AKA. This scheme uses a password associated with fingerprint and PKC to achieve the authentication between the UE and HSS.

In our scheme, we try to maintain the problems defined in the previous related works. Our proposed scheme does not need to change the original framework and the infrastructure of the LTE-A network, although a ciphered IMSI is transmitted. The authentication procedure is performed by the HSS to authenticate the UEs and the MME; therefore, there is no possibility for any occurrence of fake MME and UEs.

3. SYSTEM MODEL

In this section, we describe the system models for the SEPS-AKA including, the network model and the threat model.

3.1. Network Model

In this subsection, we will explain the LTE-A network architecture and the original authentication and key agreement protocol used in the LTE-A network

3.1.1 LTE-A Network Architecture

The architecture of the LTE-A network is mainly composed of two components as depicted in Figure 1; the Evolved Packet Core (EPC) and Evolved Universal Terrestrial Radio Access Network (EUTRAN). [2][16].

The EPC represents the wired part in the network, which is responsible for the overall control of UEs and the bearer establishment. Each entity in the EPC has a responsibility as follows;

- MME to manage bearer and connection.
- HSS to maintain the user subscription data and MME identities.
- Packet Data Network (PDN) and Packet Data Network Gateway (PGW) to perform mobility anchor and internetworking within the 3GPP and non-3GPP technologies respectively.

- Policy Control and Charging Rule function (PCRF) to control decision making of flow and Quality of Service (QoS).

EUTRAN is the Radio Access Network (RAN) in the LTE-A system; mainly two components are included [2][16];

- UE which is the mobile phone handset.
- Evolved Node B (eNB) presents the Base Transceiver System (BTS) and the Base Station Subsystem (BSS) in the non-3GPP technologies.

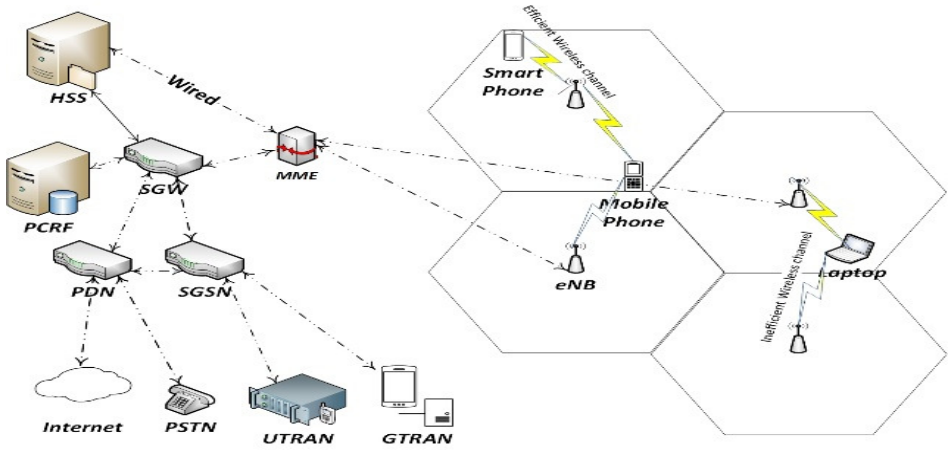


Figure 1: LTE-A Network Model

Number of eNBs is interconnected via the EUTRAN to manage multiple cells and to allow the interaction between different protocol layers in order to perform the radio resource management, header compression, securely capabilities and connectivity functions.

The security of cellular communication is a very important issue to subscribers. Attacker can exploit any security flaw to perform their goals. Telecommunication, video and audio streaming, mobile banking, data transmission, commences and etc. are a mobile application that may be attacked and hence causes a sophisticated problems. Authentication is the security issue that verifies users to the network. The problem in this issue allows unauthorized users to access the network.

Consider the system and the communication models of the LTE-A as depicted in Figure 1, the MMEs are the connection link between the HSS and the UEs. In our propose scheme, we do not introduce any modification on the original infrastructure of the LTE-A. Only we introduce a soft modification on the parameters and algorithms that are adapted in the system entities; such as the RSA scheme in the UEs and the HSS, to generate and regenerate parameters, and the certificate authority, to distribute PKs to between MMEs and UEs [2][16].

3.1.2. LTE-A Authentication Procedure

EPS-AKA authentication procedure was proposed in the 3GPP release 9 for LTE networks. EPS-AKA can broadly be divided into two stages: (1) authentication data distribution, and (2) user authentication and key agreement. The former enables the home network (HN) of a mobile equipment (ME) to distribute authentication data to the serving network (SN) where the ME device is visiting. The latter is to establish new session keys between the ME and the SN. The EPS-AKA protocol works as follows [10]:

a) An UE sends an access request message to the MME.
b) Upon receiving a request, the MME launches an authentication procedure by asking the UEs identity (IMSI).
c) In response to the MME, the UE sends its identity (IMSI).
d) The MME sends an authentication data request message containing IMSI to the HSS for acquiring Authentication Vectors (AVs).
e) The HSS first generates AVs for the MME, an AV comprising a RAND, XRES, AUTN and KASME in steading of IK and CK in UMTS AV. The AV is expressed as AV = RAND||XRES||KASME||AUTN, and AUTN = SQN $\oplus$(AK || AMF || MAC), where $\oplus$ is a simple bitwise XOR and || is a simple concatenation operations.
f) The HSS sends back an authentication data request message including the generated AV (for the corresponding UE), so that the MME is authorized to authenticate the requesting UE.
g) Upon receipt of authentication vectors, the MME sends RAND and AUTN piggy-backed on authentication request to the UE, enabling the ME to verify the correctness of SQN and compute the RES.
h) The UE verifies the correctness of SQN by computing MAC and comparing it with the MAC carried in AUTN. If matched, the ME computes and sends the corresponding response RES back to the MME in an authentication response message.
i) Once the MME receives and verifies RES correctly, it chooses the corresponding KASME as the session key to protect its communication with the ME. In addition, the ME calculates its KASME accordingly.

3.2. Threat and Trust Model

The violation of the wireless network systems is a common target of hackers, thus, in this subsection; we consider two type of attacks that may violate the security of the LTE-A system such as the cyber-attack and the side channel attack [12].

Cyber-attack is any type of offensive maneuver employed by hackers that targets computer information systems, infrastructures, and computer networks by various means of malicious actions usually originating from an anonymous source that steals, alters, or destroys a specified target by hacking into a susceptible system. Cyber-attacks can range from installing spyware on computer systems to attempts to destroy the infrastructure of entire nations. Cyber-attacks have become increasingly sophisticated. In the LTE-A system a cyber-attack may be employed as a malicious such as, MME, UE, Home eNB (HeNB), and non-3GPP access point, in order to break down the system [12].

A side channel attack may be violate the security of the LTE-A system, since it relies on the relationship between information emitted through the side-channel and the secret data depending on information gained from the physical implementation of a cryptosystem. In the LTE-A network, femto and micro cells are good environments for side channel attack to be efficient where a technical knowledge of the internal operation of the system and powerful statistical methods are defined [12].

The trusted model is a TTP in the PKI that creates the public/private keys [12]. However, a meaningful trust model for a PKI must consider the semantic assumption and human cognition of trust relationship, such as the legal constricted agreements between participants and how identity information is displayed and represented. In our proposed scheme, we use the Pretty Good Privacy (PGP) to authenticate UE to the HSS. PGP [12] is a free version commercial encryption entity used to authenticate parities, the PGP allows every user to play a role of relaying parity where each one can sent certificate to each other. Therefore, PGP defines three methods for users

and relaying party to obtain a public key of other users; 1) a secure out-of-band channel, such as physical meeting, 2) online trust decision based on introductions of new certificate from previously trusted users, and 3) based on discretionary trust decision when receiving a public key [12].

4. PRELIMINARIES

In this section, both the public key infrastructure and the RSA scheme are presented as our preliminaries, since the SEPS-AKA is based on both of them.

4.1. Public Key (PK) Cryptography

It is asymmetric cryptography involving the use of two separate keys, unlike symmetric encryption, that uses only one key [13][14]. The use of two keys has profound consequences in the areas of confidentiality, key distribution, and authentication. Public-Key Cryptography was developed to address two key issues. First, key distribution, in how to have secure communications in general without having to trust a KDC with your key. Second, digital signatures, in how to verify a message comes intact from the claimed sender. Public-key cryptography involves the use of two keys. First, Public-key is known to everybody in order to use for encrypting messages, and verifying signatures. Second, Private-key, known only to the recipient, used to decrypt messages, and sign (create) signatures. Public key cryptography applications are classified into three categories: 1) Encryption/decryption; the sender encrypts a message with the recipient's public key, 2) Digital signature; the sender signs a message with its private key, either to the whole message or to a small block of data that is a function of the message, and 3) Key exchange, two sides cooperate to exchange a session key. The main advantage of public key cryptography is the asymmetricity since who encrypts message or verifies signature cannot decrypt same messages or create same signatures, therefore, it is infeasible to determine private key from public [13][14].

4.2. The RSA Scheme

The security principle of the RSA scheme is based on hardness of the factorization problem due to the cost of factorizes large numbers [13][14]. Each user generates a public/private key pair by selecting two large prime numbers at random: p, q. compute $n = p * q$ and $\phi(n)=(p-1)(q-1)$. Randomly, the RSA scheme selects an odd number e, where $1< e < \phi(n)$, $\gcd(e , \phi(n)) =1$, $e*d=1 \bmod \phi(n)$, and $0\leq d \leq n$. Then publishes their public key $PU=\{e,n\}$ and keeps secret private key $PR=\{d,n\}$. To encrypt a message, M, the sender obtains a public key of the recipient and computes the cipher text, $C = M^e \bmod n$, where $0 \leq M < n$. To decrypt the cipher text C, the owner of the message uses their private key, $PR=\{d,n\}$, and computes $M = C^d \bmod n$ [13][14].

5. PROPOSED SCHEME (SEPS-AKA)

Figure 2 describes the SEPS-AKA proposed scheme, the workflow of the SEPS-AKA is the similar to the framework of the original EPS-AKA scheme. The methodology of our proposed system uses the two methods explained in the previous section to enhance and adapt the privacy of the original LTE-A authentication procedure. First, the infrastructure of the public key cryptography is used to encrypt the exchanged data between LTE-A network entities. Second, the RSA scheme computation is used to compute the used parameters in the previous section. A pre-shared secret key K is used as the original LTE-A network where the key was industrially preset to the devices and stored physically in the USIM and to the HSS. The workflow of the SEPS-AKA scheme is described as follows:

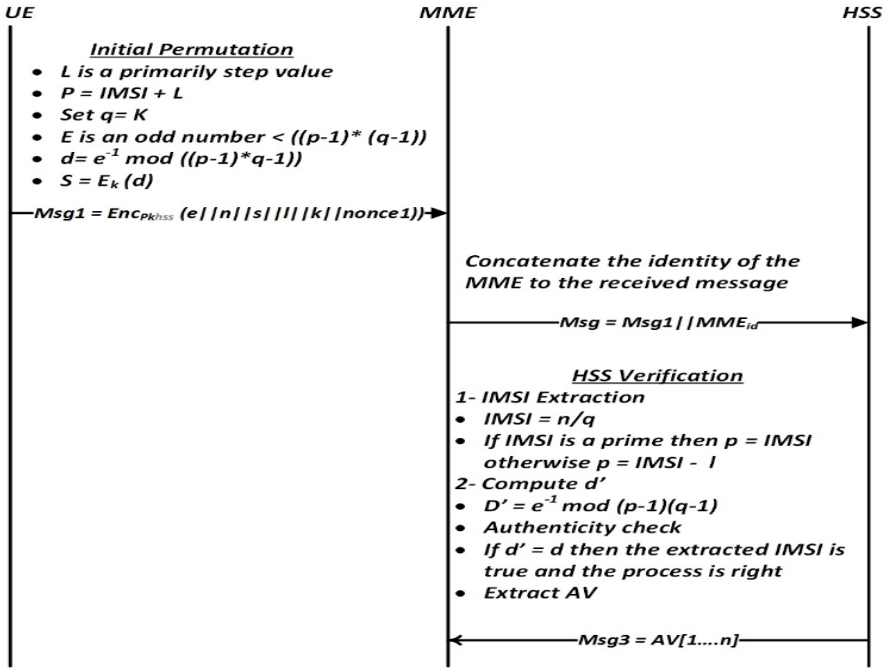


Figure 2: The SEPS-AKA Scheme

5.1. UE ⟶ MME

In this stage, UE initiates five parameters; *p, q, n, d*, and *L*. these parameters are computed based on the RSA scheme computations using the IMSI and k as follows:

a) Parameter Initiation

○ Select large prime number p.

$$p = IMSI + L \quad (1)$$

Where *L* is an integer number computed when the IMSI is not prime as L = 0, if IMSI is prime, otherwise L is the step value of the next prime after IMSI.

○ Set q as a random large prime number

○ Compute *n* :

$$n = p * q \quad (2)$$

○ Compute large number d:

$$d = e^{-1} \bmod (p-1)(q-1) \quad (3)$$

Where *e* is an odd number < *((p-1) *(q-1))*

b) Parameter Encryption: Encrypts d using a standard encryption algorithm (AES) to provide an encrypted parameter, *s*.

$$s = E_k (d) \quad (4)$$

c) Message Building and originating: Concatenates the last parameters p, n, q, s, L and a nonce1 together, encrypts the result using the public key of the HSS (PK_{HSS}) and originating the encrypted message to the HSS.

$$msg1 = E_{PK_{HSS}}(e \,||\, n \,||\, s \,||\, q \,||\, L \,||\, nonce1) \quad (5)$$

5.2. MME ⟶ HSS

Once MME receives msg1, it builds msg2 as illustrated in (6), which is the received message from the UE msg1 plus its identity MME_{id} and originates the message to the HSS

$$msg2 = msg1 \,||\, MME_{id} \quad (6)$$

Where MME_{id} is the identity of the MME

5.3. HSS ⟶ MME

Once the HSS receives *msg2* from the MME, it decrypts the message and executes the following:

- **IMSI extraction**

$$IMSI = (n / q) \quad (7)$$

- Check the primarility of the extracted IMSI,
  If IMSI is prime then

$$p = IMSI \quad (8)$$

Otherwise

$$p = IMSI - L \quad (9)$$

Where L is an integer number transmitted from the UE, and was computed when the IMSI is not prime in equation (1) as; $L = 0$, if IMSI is prime, otherwise L is the step value of the next prime after IMSI

- **IMSI Verification:**

Compute d' as illustrated in (10); a large number computed based on the RSA scheme computation, in order to verify the IMSI in the HSS side. The IMSI verification in the HSS is done as follows; check $d' = d$, if true then the IMSI is true; this process is not supported in the original EPS-AKA scheme, while the IMSI is verified in the HSS using database query.

$$d' = (e^{-1}) \bmod (p-1)(q-1) \quad (10)$$

After the IMSI verification is done correctly the remained steps are doing normally as the original EPS-AKA scheme.

6. SECURITY AND PRIVACY ANALYSIS

In this section, we analyze the security of our scheme to demonstrate that it meets the security requirements of the LTE-A systems. In our scheme, three levels of security are used: PKI, the RSA scheme computations and nonces. Nonces are random numbers generated by UE, MME and HSS to use in generating challenge messages toward the opposite side. A different Nonces are used in each authentication procedure, therefore, the reusing of these Nonces are not efficient. An out-of-sync situation will lead to authentication failure.

Consider a cyber-attack, in which a malicious UE aims to register to the LTE-A network, the malicious UE need to gain the computation of the RSA scheme parameters (p, q, s, and n) and to

gain the encryption information, which is required to prepare the message before sending from the UE. For the malicious MME, at the first registration time, the MME is considered as a gateway to route encrypted messages from the UE to HSS, while the MME must concatenate its certificate with the routed message in order to prove its authenticity to the HSS. Therefore, we consider the presence of the cyber-attack is impossible.

Consider a legal UE is worked through a femto and micro cells, which are two authorized environments uncontrolled by the LTE-A network, IMSI is not sent through the authentication message but is computed by the RSA scheme parameters (p and q), which is an NP problem, while the side channel attacks need a technical knowledge of the internal operation of the system and powerful statistical methods to be efficient.

In addition, our scheme prevent replay, impersonate attacks, the Man in the Middle, and DoS attacks. The replay attack is prevented by using the nonces in the transmitted messages, therefore, it is no possibility to use this message again. In addition of using the PK cryptography to encrypts the transmitted messages, the IMSI, the legal identity of the UE, is not transmitted in clear text over the transmitted messages, therefore, the attack cannot able to impersonate the identity of the UE. The MitM and DoS attacks are prevented as; if a member is able to sniff PKI, it still cannot computes the IMSI using the RSA scheme computations, although these messages are sent with PKI protection therefore, the attacker may not be able to hack this data since PKI is having the residency of DoS and MitM attacks. Therefore, the SEPS-AKA scheme attained the security requirements such as privacy, confidentiality, authentication, and data integrity.

Table 1. EPS-AKA Security Requirements

	Entity mutual authentication	Privacy	Confidentiality	Data Integrity
SEPS-AKA	Yes	Yes	Yes	Yes
EC AKA	Yes	Yes	Yes	No
SP AKA	Yes	No	Yes	No
HSK AKA	Yes	No	No	No
EPS-AKA	Yes	No	No	No

Table 1 is a comparison between SEPS-AKA scheme and the other previous EPS-AKA schemes in addition to the original EPS-AKA. The SEPS-AKA scheme adopts the same secured architecture as the EPS-AKA protocol. Therefore, it has the same security threshold in most situations. As illustrated in Table 1, the SEPS-AKA scheme can attain the security requirements as follows:

6.1. Entity mutual authentication

All schemes attain the entity mutual authentication since a UE is identified to the HSS by its IMSI. However, comparing with the SEPS-AKA scheme, where an UE is identified mathematically by its IMSI, as mentioned in equations (1), (7), (8), and (9), and the mathematical computation were based on the RSA scheme. The original and the EC-AKA schemes, the transmission of the IMSI gains the probability to different previous attack. In the SP-AKA and the HSK-AKA, the first step is begin from the MME that maintain a high probability of the presence of cyber-attack and the side channel attack.

6.2 Privacy

To ensure user privacy, the IMSI should be confidentiality protected. It should never be transmitted without protection. The EC-AKA achieves the privacy since an encrypted IMSI is transmitted. But the remainder schemes has no privacy since the IMSI is transmitted in clear text. But the SEPS-AKA scheme attains a high level of privacy where the IMSI is protected using the public key of the HSS.

6.3 Confidentiality

Confidentiality includes cipher algorithm agreement, cipher key agreement, confidentiality of user data and confidentiality of signaling data. The SEPS-AKA scheme follows the mechanism of the EPS-AKA protocol and hence is successful with these demands.

6.4 Data Integrity

Data integrity includes integrity algorithm agreement, integrity key agreement, data integrity and original authentication of signaling data. As illustrated in table I, No one of the scheme presented in table 1 except the SEPS-AKA scheme attain these purposes since no one of these schemes provide a level of verification of the IMSI, while in the SEPS-AKA scheme a high level of IMSI verification is performed using the IMSI extraction as illustrated in equation (9).

7. PERFORMANCE EVALUATION

The evaluation of the performance of the SEPS-AKA scheme is compared to HSK-AKA [7], EC-AKA [8], SP-AKA [9] and the original EPS-AKA. Therefore, two comparison criteria's will be discussed; bandwidth consumption and computation overhead.

7.1.Bandwidth consumption

Measuring the bandwidth consumption requires defining the employed cryptographic algorithm. Suppose the RSA scheme with 1024-bit key, therefore, the measuring of cipher text size as following [14]:

o Compute n:

$$n = \sum plaintext\ length\ in\ bytes \tag{11}$$

Where n is the length of the transmitted plaintext in bytes.

o Divides plaintext into equal blocks (16 byte)

$$s = ceil(\frac{n}{16}) \tag{12}$$

Where *S* is the integer number of blocks of the plaintext.

o Compute Ciphertext length c_len as :

$$c\_len = \left(1 + \frac{floor(RSAkeysize-1)}{8}\right) * s \tag{13}$$

Where S is the number of blocks

As shown in the table 2 and figure 3, the SEPS-AKA scheme consumes a bandwidth less than the EC-AKA, but comparing with SP-AKA, HSK-AKA and original EPS-AKA schemes, SEPS-AKA consumes a greater bandwidth. Since SP-AKA is not follow the same framework of the original EPS-AKA. HSK-AKA secures the original AKA based on the pseudonyms approach without PKI, and there is no security consideration on the original EPS-AKA

Table 2. Bandwidth Consumption

Scheme	Total bandwidth in bytes
SEPS-AKA	676
EC-AKA	816
SP-AKA	240
HSK-AKA	336
EPS-AKA	276

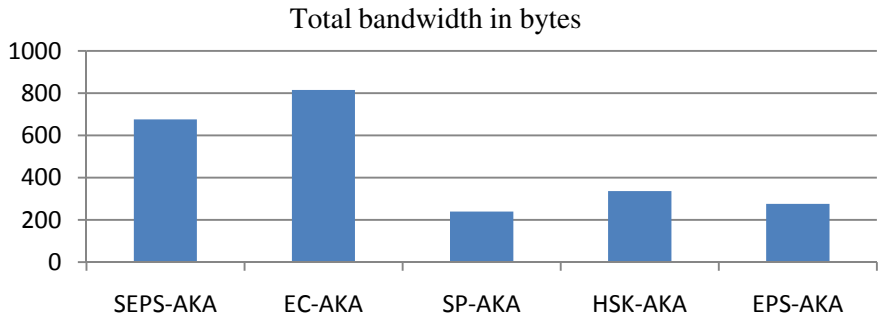


Figure 3: Bandwidth Consumption

7.2.Computation Overhead

To measure the computational overhead, we use crypto++5.6.0 benchmark which is compiled with Microsoft visual C++ 2005 SP1 and runs on Intel core 2 1.83 GHz processor under WINDOWS VISTA in 32 bit mode [15].

As illustrated in table 3 and figure 4, SEPS-AKA preserves the computational overhead compared to the EC-AKA scheme, while SP-AKA scheme provide less than computational overhead since, it is not recognized based on the original framework of the original EPS-AKA scheme. In addition, the original EPS-AKA and HSK-AKA schemes have no security aspects, therefore, there is no computational overhead attached with them.

Table 3. Computational overhead

Scheme	Computational overhead in microsecond
SEPS-AKA	39540
EC-AKA	89540
SP-AKA	2926

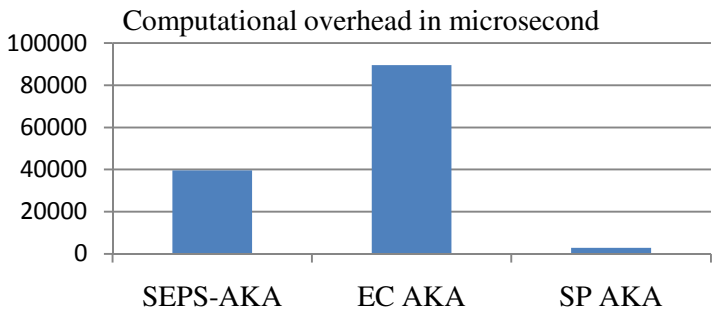


Figure 4: Computational Overhead

8. CONCLUSION

In this paper, we have proposed a secure and efficient EPS-AKA scheme, SEPS-AKA, using PKI and the RSA scheme computations in order to maintain the problems in the LTE-A authentication and key management. Compared with other authentication protocols, our proposed scheme robustly achieves security requirements including; privacy, authentication, confidentiality, and data integrity. Moreover, as the major contributions of the paper, extensive security analysis shows that the SEPS-AKA scheme is secure against various malicious attacks such as cyber and side channel attacks. Furthermore, the SEPS-AKA scheme has a high withstanding to the replay, DoS, MitM, and Impersonation attacks. The performance evaluation shows that the SEPS-AKA scheme achieves good bandwidth consumption and less computation overhead.

ACKNOWLEDGEMENTS

This paper is supported in part by the Zamalah Fellowship Program, Gaza, Palestine.

REFERENCES

[1] P. Bhat, S. Nagata, L. Campoy, I. Berberana, T. Derham, G. Liu, X. Shen, P. Zong, and J. Yang, "LTE-advanced: an operator perspective, "Communications Magazine, IEEE, vol. 50, no. 2, pp. 104–114, February 2012.

[2] J. Cao, M. Ma, H. Li, Y. Zhang, and Z. Luo, "A survey on security aspects for LTE and LTE-A networks," Communications Surveys Tutorials, IEEE, vol. 16, no. 1, pp. 283–302, First 2013.

[3] S. Kanchi, S. Sandilya, D. Bhosale, A. Pitkar, and M. Gondhalekar, "Overview of LTE-A technology," in Global High Tech Congress on Electronics (GHTCE), 2013 IEEE, Nov 2013, pp. 195–200.

[4] J.-K. Tsay and S. F. Mjlsnes, "Computational security analysis of the UMTS and LTE authentication and key agreement protocols," Report arXiv: 1203.3866v2, Norwegian University of Sciences and Technology (NTNU), Department of Telematics, Norway, 2013.

[5] Y. Park and T. Park, "A survey of security threats on 4g networks," in Globecom Workshops, 2007 IEEE, Nov 2007, pp. 1–6.

[6] M. Purkhiabani and A. Salahi, "Enhanced authentication and key agreement procedure of next generation evolved mobile networks," in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, May 2011, pp. 557–563.

[7] K. Hamandi, I. Sarji, A. Chehab, I. Elhajj, and A. Kayssi, "Privacy enhanced and computationally efficient HSK-AKA LTE scheme," in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, March 2013, pp. 929–934.

[8] J. Abdo, J. Demerjian, H. Chaouchi, and G. Pujolle, "EC-AKA2 a revolutionary aka protocol," in Computer Applications Technology (ICCAT), 2013 International Conference on, Jan 2013, pp. 1–6.

[9] J. B. Abdo, J. Demerjian, K. Ahmad, H. Chaouchi, and G. Pujolle, "EPS mutual authentication and crypt-analyzing SP-AKA," in Computing, Management and Telecommunications (ComManTel), 2013 International Conference on, Jan 2013, pp. 303–308.
[10] C. Lai, H. Li, R. Lu, and X. S. Shen, "SE-AKA: A secure and efficient group authentication and key agreement protocol for LTE networks," in Computer Networks, Sep 2013, pp. 3492 – 3510.
[11] Y. Zheng, D. He, X. Tang, and H. Wang, "AKA and authorization scheme for 4G mobile networks based on trusted mobile platform," in Information, Communications and Signal Processing, 2005 Fifth International Conference on, 2005, pp. 976–980.
[12] C. Tang, D. Naumann, and S. Wetzel, "Analysis of authentication and key establishment in inter-generational mobile telephony," in High Performance Computing and Communications 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC·EUC), pp. 1605–1614.
[13] Eli, Atilla and R. Shankaran, "Theory and practice of cryptography solutions for secure information systems," in IGI Global, Sep 2013, pp. 1–351.
[14] W. Stallings, "Cryptography and network security: Principles and practice," in Prentice Hall, Jan 2011, p. Fifth Edition.
[15] W. Dai, "Crypto++ 5.6.0 benchmarks," MARCH 2009.
[16] P. Rengaraju, C.-H. Lung, and A. Srinivasan, "Measuring and analyzing WIMAX security and QoS in testbed experiments," in Communications (ICC), 2011 IEEE International Conference on, June 2011, pp. 1–5.

AUTHORS

Zaher Jabr Haddad received the BSc degree from the computer science department, faculty of applied science, Al-Aqsa University, Palestine in 1999 and MSc degree from the Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt in 2007. He is currently working toward the Ph.D. degree in Information Technology Department, Faculty of Computers and Information, Cairo University, Egypt. His interest in wireless network security and LTE networks

Sanaa Taha received the BSc and MSc degrees from the Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt, 2001, 2005, respectively, and the PhD degree in the Electrical and Computer Engineering from the University of Waterloo, Canada in 2013. She is currently an associated professor in the Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt. Her research interest include wireless network security, mobile networks security, mobile management, and applied cryptography.

Imane Aly Saroit Ismail received the B.Sc, MSc, and PhD from Department of Communication, Faculty of Engineering, Cairo University, Egypt, in 1985, 1990, and 1994, respectively. She is currently a full professor in the Department of Information Technology, Faculty of Computers and Information, Cairo University, Egypt. Her research interest include wireless network security, mobile networks security, mobile management, and applied cryptography.

SECURITY ANALYSIS ON PASSWORD AUTHENTICATION SYSTEM OF WEB PORTAL

Heekyeong Noh[1], Changkuk Choi[2], Minsu Park[3], Jaeki Kim[4], Seungjoo Kim[5]

CIST (Center for Information Security Technologies),
Korea University, Seoul, Korea
[1]hknoh@korea.ac.kr, [2]nikojin@gmail.com, [3] minsoon2@korea.ac.kr,
[4] jack2@korea.ac.kr, 5skim71@korea.ac.kr

ABSTRACT

Portal site is not only providing search engine and e-mail service but also various services including blog, news, shopping, and others. The fact that average number of daily login for Korean portal site Naver is reaching 300 million suggests that many people are using portal sites. One of the most famous social network service, Facebook subscribers to reach 1.2 billion 30 million people at the time of the February 2014. With the increase in number of users followed by the diversity in types of services provided by portal sites and SNS, the attack is also increasing. Therefore, the objective of this study lies in analysing whole procedure of password authentication system of portal sites, SNS and analysing the security threat that may occur accordingly. Also, the security requirement corresponding to analysed security threat was extracted and the analysis on implementation of security requirements by portal sites and SNS was conducted.

KEYWORDS

Password Authentication System of Web Sites, Threat of Web Sites, Security Requirement of Web Sites, Attack Potential of Password Systems

1. INTRODUCTION

The dictionary definition of portal is 'entrance' or 'gateway' and the term portal site (hereafter referred to as a portal) signifies a site which plays the role of a gateway by collecting and organizing enormous quantities of internet data so that users can easily access the particular data they require. Although the original format of portals was primarily based around search engines and e-mail services, they currently provide widely varied web services such as those related to news, shopping, and blogging. Furthermore, the e-mail accounts provided by portals are used as IDs for social network services (SNSs), such as Facebook and Twitter, and other web services and applications, and even as a way to authenticate users who have forgotten their account passwords. As such, portal accounts are increasingly used not merely for e-mail communication but are connected to services providing a wide range of web-based activities.

When a portal account is used at another web service or portal, the security strength of both services decreases to that of the site with the weakest security based on the principle of

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 71–92, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41207

minimization, 'since the security is entwined in a chain, the weakest security strength determines the security strength of the whole' [1]. Such a chain occurs, for example, when a user creates a Google (www.google.com) account and provides his Naver (www.naver.com) e-mail account address as personal information. A Google account requires a minimum password length of 8 characters but does not require a combination of numbers and upper and lower case letters. Naver, on the other hand, requires a minimum password length of only 6 characters and also does not require a combination of numbers and upper and lower case letters. However, the weaker password requirements of Naver accounts reduce not only the security of Naver accounts, but also of any Google account created using a Naver one. Thus, the security of the Google account is reduced to that of the Naver one, since both accounts can be hacked after obtaining the Naver password, assuming the IDs of the accounts are named the same or the linked account is known by some other means. To hack into the Google account, the attacker can simply request a password reset and request user authentication through the Naver account. At the moment, although the Naver account password is easier to obtain using a complete enumeration survey than the Google one, it is trivial for an attacker to access the Google account after acquiring the Naver one.

Therefore, an analysis of portal sites' authentication systems at member registration, login, password reset step 1, and password reset step 2, including via SNS, were conducted in this study. Security threats that may exist in the authentication procedure of each portal and SNS, and the security countermeasures against such threats were clarified. Afterwards, a quantitative analysis of attack threats and the implementation by portals of their corresponding countermeasures were conducted by applying a standardized set of security criteria to each portal.

2. RELATED WORKS

2.1. Password Authentication Systems

As online services and applications become more sophisticated, users are increasingly required to create an account to receive a service. Studies of security of authentication systems undertaken until now mainly focus on the security vulnerabilities of ID-password based authentication systems and their countermeasures. When users create distinct accounts for a different variety of services they often use the same IDs and passwords due to memory limitations and the inconvenience of managing multiple accounts differently. In such cases, there is the problem that an attacker can access the user's other accounts by just acquiring the authentication information of one [2]. Furthermore, if the account obtained is just a portal, the scale of local damage may be small, but severe damage may result when the same user's accounts at sites related to banking and payment, including internet banking and internet shopping malls, amongst others, are also obtained. Although an answer to a security question may be requested from users during the user authentication procedure for password resets if a user forgets his password, the answers to such questions can often be either too easily guessed or too difficult for even the user himself to remember. Therefore, in order to resolve such problems, studies were undertaken to improve the security of the design and selection of account registration security questions [3]. Afterwards, in order to resolve the problem of remembering several IDs, many service providers started utilizing the most frequently used e-mail address as ID. Recently, the analysis of the security threat to users' accounts and privacy was conducted in relation to such elements as the password management plans of service providers and multiple uses of the same password, and possible solutions to their weaknesses proposed [4].

Studies analysing the security of various password authentication systems have been conducted and they recommended security-hardening methods. These included CAPTCHA, after consecutive login attempts, that confirm that the login device is human-operated by requiring

only human-discernible answers, salting techniques which use random numbers for the application of a password into a hash function so the password can be safely encrypted, and key strength algorithms which respond to consecutive attempts by lengthening decoding times by repeated encryption of the password with a hash function. Also, an algorithm to examine the security strength of passwords and effectively perform safer password creation was proposed. Entropy-based security assessment is available and entropy was first proposed as a concept for measuring the uncertainty and randomness for security by Claude Shannon [5]. In order to measure the entropy value of a password, the distribution of password length, text placement, numbers of letter types, and contents of the text are set as standards and the sum of all their entropies is the total entropy value [6]. Then, the method to quantitatively examine the security of passwords was proposed by the creation of a PQI (password quality indicator) which measures the security strength of passwords by considering their entropy [7].

Moreover, various methods to promptly and accurately determine password security strength were proposed to help users create strongly secure passwords. It was proven that a way to prevent successful pre-emptive attacks can be first sorting passwords into those of high and low security strength by conducting pattern analysis [8]. Using this method, users are prevented from selecting easily guessed passwords at password creation. The security strength of passwords was enhanced by comparing the password entered by the user during the password creation procedure against a list of those from dictionaries used by attackers in their pre-emptive attacks and disallowing any matching passwords [9].

However, the limitation of previous studies is that they focus on and propose countermeasures for the problems of password and ID based authentication by service providers only rather than the security threats or attacks on the password authentication system as a whole. The studies focused on the security of passwords rather than the password authentication system as a whole. Therefore, the objective of this study is to analyse the whole portal password authentication system including member registration, login, password reset step 1, and password reset step 2, conduct analysis on the security vulnerabilities of the overall system, and determine the security requirements for the countermeasures against those vulnerabilities.

2.2. Attack Potential of Common Evaluation Methodology (CEM)

Attack potential refers to a function of expertise, resource and motivation presented by Criteria Evaluation Methodology (CEM) in the CC, which consists of elapsed time, expertise, knowledge about a target of attack, period of easy exposure to attack and equipment and quantitatively shows attack potential of the target of attack by giving values to each element [10].

2.2.1. Elapsed time

Elapsed time is the total amount of time taken by an attacker to identify that a particular potential vulnerability may exist in the TOE, to develop an attack method and to sustain effort required to mount the attack against the TOE.

2.2.2. Expertise

Expertise refers to the level of generic knowledge of the underlying principles, product type or attack methods. The identified levels are as follows:

- Layman: unknowledgeable compared to experts or proficient persons, with no particular expertise

- Proficient: knowledgeable in that they are familiar with the password attack tools and methods
- Expert: familiar with implementing in password attack tools, operation algorithm of password authentication systems.

2.2.3. Knowledge of target of attack

Knowledge of the TOE refers to specific expertise in relation to the TOE. This is distinct from generic expertise, but not unrelated to it. Identified levels are as follows:

- Public: information gained from the Internet
- Restricted: knowledge that is controlled within the developer organisation and shared with other organisations under a non-disclosure agreement
- Sensitive: knowledge that is shared between discreet teams within the developer organisation, access to which is constrained only to members of the specified teams
- Critical: knowledge that is known by only a few individuals, access to which is very tightly controlled on a strict need to know basis and individual undertaking

2.2.4. Period of easy exposure to attack

Period (chance) related to elapsed time, when an attacker can approach the target of attack.

- Unnecessary/unlimited access: the attack doesn't need any kind of opportunity to be realised because there is no risk of being detected during access to the TOE.
- Easy: access is required for less than a day
- Moderate: access is required for less than a month
- Difficult: access is required for at least a month

2.2.5. Equipment

Equipment refers to the equipment required to identify or exploit a vulnerability [11].

- Standard equipment is readily available to the attacker, either for the identification of a vulnerability or for an attack.
- Specialised equipment is not readily available to the attacker, but could be acquired without undue effort.
- Bespoke equipment is not readily available to the public as it may need to be specially produced.
- Multiple Bespoke is introduced to allow for a situation, where different types of bespoke equipment are required for distinct steps of an attack.

Table 1. Password Attack Tools

Tool	Equipment Level
Cain and Abel	Standard
John the Ripper	Standard
SolarWinds	Standard
RainbowCrack	Standard
wfuzz	Standard
Medusa	Standard
THC Hydra	Standard

Table 2 identifies the factors discussed in the previous and associates numeric values with the total value of each factor.

Table 2. Attack Potential of Common Criteria

Factor		Value
Elapsed time	≤ 1 hour	1
	≤ 1 day	3
	≤ 1 week	5
	≤ 1 month	7
	≤ 6 month	10
	> 6 month	15
Expertise	Layman	0
	Proficient	3
	Expert	6
	Multiple expert	8
Knowledge about target of attack	Public	0
	Restricted	3
	Sensitive	7
	Critical	11
Access to object	Non-Restricted	0
	Easy	1
	Normal	4
	Hard	10
	None	*
Equipment	None	0
	Standard	4
	Bespoke	7
	Multi bespoke	9

3. ANALYSIS ON AUTHENTICATION SYSTEM

This chapter conducts analysis on authentication system of portals, SNS and examines the improvement plan for authentication system. The analysis subjects of this study include Naver (www.naver.com), Nate (www.nate.com), and Daum (www.daum.net) for Korean portals and Google (www.google.com), Yahoo (www.yahoo.com), and MSN (www.msn.com) for U.S.s portals and Facebook (www.facebook.com), Twitter (www.twitter.com) for SNS. Also, authentication procedure of portal was divided into 4 steps of member registration, login, password reset-phase 1, and password reset- phase 2.

3.1. Member Registration

In order to prevent random account creation with automated registration programs, portals have been developing CAPTCHAs, mobile phone authentication and e-mail authentication. Google requires the input of a CAPTCHA without provision of e-mail and mobile phone authentication as the collection of e-mail addresses and mobile phone numbers is optional. However, if users fail to enter the CAPTCHA, mobile phone authentication is provided. MSN requires the input of a CAPTCHA and Yahoo only requests mobile phone authentication without CAPTCHA input. All

Korean portals including Naver, Nate, and Daum do not require the input of a CAPTCHA and prevent automatic registration with mobile phone or e-mail authentication. In the case of Korean portals, the number of IDs issuable to a single mobile phone number is limited to 3 to prevent random account creation. However, due to a policy restricting the number of IDs issued, attackers are stealing currently used accounts for malicious use. Contrastingly, Facebook and Twitter do not provide any system preventing automatic registration so attackers abuse the service by sending SPAM messages advertising illegal content to normal users.

Also, portals provide an overseas IP block service to block attack attempts from overseas. Provision and setting of the overseas IP address block service is under user control. Table 3 illustrates the result of a survey of the basic service provided by each company when the user has not set the overseas IP block service. Google, Naver, and Nate allow login after confirming the user through personal information, e-mail, or phone number authentication once a login attempt from an overseas IP is detected. Then, an alert e-mail is sent in order to inform users about the detection of login attempts from an overseas IP address. Yahoo and Daum allow logins without a separate user authentication procedure and send an e-mail alert to the user. However, MSN allows login with neither a user authentication procedure nor an alert e-mail.

Table 3. Prevention services of portal from Automatic registration

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
CAPTCHA	-	-	-	○	○	-	-	-
Email	-	◐	◐	-	-	-	-	-
Mobile Phone	○	◐	◐	-	-	○	-	-

3.2. Login Attempts

3.2.1. IP Address Security

Portals and SNS shall provide overseas IP block service in order to correspond to overseas attack attempts. Provision and setting of oversea IP address block service follows the decision of user. Table 4 illustrates the result of conducting survey on basic service provided by each company in case user did not set overseas IP block service. Google, Naver, and Nate allow login after confirming the user through personal information, e-mail, or phone number authentication once the login attempt with overseas IP is detected. Then, alert e-mail is sent in order to inform users about the detection of login attempts with overseas IP address. Facebook allows login after confirming the user through personal information. Yahoo and Daum allow login without separate user authentication procedure and send alert e-mail to the user. MSN allows login without user authentication procedure and alert e-mail.

Table 4. Status of Overseas IP protection services

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Authentication	○	○	-	○	-	-	○	-
Email Alarm	○	○	○	○	-	○	-	-

Below Fig.1 is the screen of Naver and Google to inform users about overseas login attempt. Naver conducts authentication with the input of name and date of birth and Google conducts authentication with the use of mobile phone authentication, e-mail authentication, and password hint & answer.

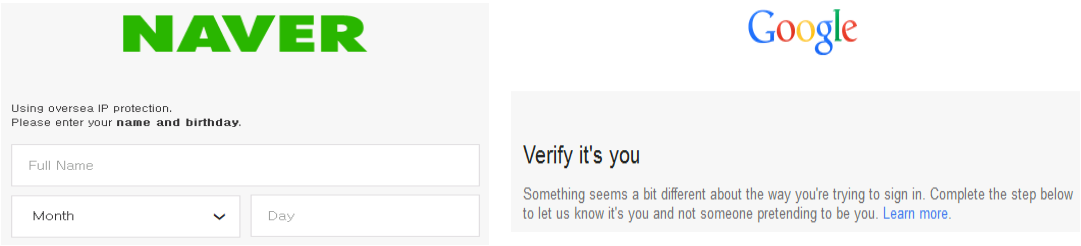


Figure 1 . Block Login attempts from overseas IP – Naver, Google

3.2.2. Consecutive Login Attempts

Attackers use bot for random login attempts to user account. All portals request the input of CAPTCHA in case of certain numbers of login failure to correspond to account hacking using the bot and request for both password and CAPTCHA when incorrect CAPTCHA value was entered. Number of login failures that requires CAPTCHA input differs according to each portal. Google requires the input of CAPTCHA with random numbers of failures and other portals request for CAPTCHA with fixed number of failures. Details on number and implementation are illustrated in Table 5.

Table 5. Threshold of account lock and CAPTCHA

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
CAPTCHA	5Times	5Times	5Times	N Times	10Times	5Times	-	-
Account Lock	-	-	○	○	○	○	○	○
Failed Count	-	-	5Times	nTimes	5Times	5Times	20Times	16Times
Lock Time	-	-	3Hours	24Hours	24Hours	12Hours	12Hours	1Hours

Also, the account shall be locked for certain period of time and additional login attempts shall be blocked with the additional login failure after the input of CAPTCHA. The account of user is protected from account hacking with the account lock of 24 hours for Google and MSN, 12 hours for Yahoo, and 3 hours for Daum. However, Naver, Nate, Facebook and Twitter do not provide account lock service. Namely, the attacker can continuously attempt for account hacking in case of Naver, Nate. Additionally, Facebook and Twitter do not request for CAPTCHA but provide account lock service. Below Fig.2 presents the alert message of Yahoo and Facebook to inform users about 12 hours account lock upon detection of consecutive login attempts.

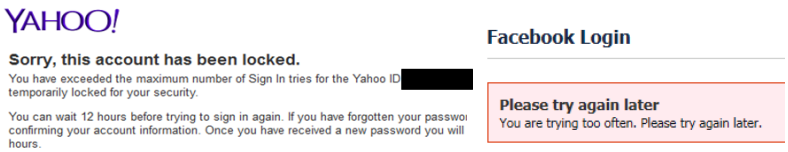


Figure 2. Lock Account – Yahoo, Facebook

3.3 Password Reset – Authentication Step 1

If a user requests a password reset, user authentication is conducted through the e-mail address registered at the time of member registration or mobile phone SMS. There are two ways to

authenticate users by e-mail address. The first method is to send the URL of the password reset webpage via e-mail. An immediate password reset is available upon login to the e-mail account registered by the user and checking the e-mail. The second method is to send an authentication code composed of numbers via e-mail. The user conducts user authentication for a password reset by entering the authentication code, given in the e-mail, on the web site. The authentication by SMS takes the same format as the second method of e-mail authentication except that the authentication code is sent via SMS. All portals and SNSs limit the number of inputs to 3 or 5 per attempt to block an enumeration survey attack on the authentication code. Also, the number of authentication code transmissions per day is restricted to 5 or 10 after which a 24 hour temporary account lock occurs.

Since Google does not require a user to enter an e-mail address or mobile phone number at member registration, that information may not be available, in which case user authentication is conducted through two channel authentication. However, in the case of portals other than Google, since they do require either a mobile phone number or an e-mail address, user authentication can more easily be conducted just through authentication step 1.

3.4. Password Reset – Authentication Phase 2

Authentication phase 2 is conducted for users who cannot access to authentication code sent via e-mail or SMS in authentication phase 1. Fig 5 is the diagram of password reset- authentication phase 2 of Korean portals and Fig 6 on the right is the diagram of procedure for U.S. portals.

In the case of Korean portals, user authentication is conducted using the resident registration number in step 2. Naver, Nate, and Daum conduct user authentication through the transmission of a copy of the identification card (or input of the Resident Registration no.), name, ID, date of birth, and sex via e-mail or fax and Nate additionally uses a method to confirm the above information via ARS. In case of US portals and Twitter, user information accumulated during account use is utilized to confirm the user in step two authentication. The information requested during the authentication step 2 of Google, MSN, Facebook and Twitter is as follows. They receive a value from the user after subdividing the information below for each step and conducting user authentication by examining the consistency between the values input and registered information. In the case of Google, the input of a contact e-mail address is requested and an e-mail including the password reset URL is sent to that address if the e-mail address input by the user matches a previously registered e-mail address, regardless of the consistency of values input afterwards. Facebook requests the answer to a security question, such as “In what city or town was your mother born?”, and if the user inputs the correct answer, then an e-mail including a password reset URL is sent.

- Other passwords used for the account
- Title of recently sent e-mail
- Folders other than default folder
- Receiver of recently sent e-mail
- Last 5 digits of prepaid card
- Name on credit card
- Date of last login
- Date of account creation
- Frequently used e-mail address
- Initial restoration e-mail address
- Last 4 digits of credit card number
- Expiration date

The difference in step two authentication of Korean and U.S. portals lies in the fact that there exists the means of authentication, resident registration no, in Korea and convenient user authentication is available accordingly thus there is no need to go through personal behaviour based user authentication procedure of U.S. portals.

4. ANALYSIS ON SECURITY THREAT AND SECURITY REQUIREMENT FOR PASSWORD AUTHENTICATION SYSTEM

The analysis on security threat that may occur during each authentication step of portals and SNS analyzed beforehand is conducted in this chapter. Analysis on possible security threat was conducted considering the threat that may occur during login procedure, password threat, and others.

4.1 Security Threat in Password Authentication System

4.1.1 Security Threat in Member Registration Stage

T1. Automatic Registration
Attackers make monetary gain through various methods such as sending SPAM mail for advertising, the distribution of malicious code to lure people to phishing sites and the posting of advertisements. Since more of these activities can be conducted if the attacker has more accounts available to him, accounts are created using automatic registration programs.

4.1.2 Security Threat in Login Stage

T2. Consecutive Login Attempts
The attacker attempts consecutive authentication using methods such as complete enumeration survey, password guessing, and others in order to obtain the password of user account. Complete enumeration survey is an attack method to obtain correct password by substituting all of possible password combinations and password guessing attack is a method to guess possible password by gathering information such as name, date of birth, family relations, and others of user. Also, there exists an attack method to attempt at authentication by substituting the information such as password that is most frequently used by the users.

T3. Phishing
The attacker outputs phishing site instead of normal site with method same as distribution of malicious code when users access to portals [12]. Since it is difficult for general users to distinguish phishing site from normal site, they input ID and password as normal and the attacker can obtain input ID and password at the moment.

T4. Keylogging
Keylogging is an attack technique which steals information by intercepting the information input with a keyboard, often using a keylogging program [13]. Although normally information input by keyboard is displayed on the monitor after processing by the OS, keylogging programs intercept the information and save it as a file as it is processed by the OS and subsequently leak that information by sending the file to a designated server. The attacker analyses the key sequences, and tries to identify those corresponding to portal logins to obtain IDs and passwords. For example, a large attack to control portal and SNS accounts using keylogging programs occurred in Dec 2013 in which about 2 million users' information was hacked from 93,000 web sites worldwide including 318,000 Facebook, 70,000 Google Gmail and 22,000 Twitter accounts, amongst others. The attacker obtains web site login records including web site IDs and passwords by installing keylogging programs on users' computers [14].

4.1.3 Password Reset- Authentication Phase 1

T5. Consecutive Login Attempts
By selecting e-mail authentication for user authentication at the password reset stage, the attacker may attempt consecutive logins to obtain the passwords of other accounts after gaining access to an email account. The difficulty of such an attack is lowered if e-mail account passwords are

weak, meaning of low entropy, facilitating the initial email account hacking [15]. In this way, the attacker can obtain the password of multiple user accounts using methods such as complete enumeration surveys amongst others.

T6. E-mail Sniffing
Sniffing refers to the tapping of others' network packets. Portals use e-mail and mobile phone authentication for user authentication at password reset step 1. E-mail authentication may send an authentication code or password reset URL to an e-mail address registered in advance, particularly at member registration. At that moment, if the attacker intercepts the e-mail sent by the portals or the password reset page through sniffing, then he can set a new password for the victim's account himself.

T7. Mobile Phone Tapping
Mobile phone tapping of an attack target is available if the attacker has installed malicious code on the victim's phone in advance. In this situation, when an SMS including the authentication code for password reset is sent to the victim, the attacker can obtain the authentication code for himself by tapping the victim's SMS. Thereby, the attacker can obtain authority over the user account by setting a new password for the victim's account.

4.1.4 Password Reset – Authentication Phase 2

T8. User Information Guessing
In the case of Google and MSN, user authentication is conducted using user account information when the user cannot use e-mail or mobile phone authentication. At the moment, information requested by portals can include the time of recent login, time of account creation, contact e-mail address and folder names. The attacker disguises himself as a target user by entering guessable information specific to the target. Particularly, when security questions are used, such as for Yahoo, the answers to the questions can be guessed when combining account information available from SNS accounts [16]. The attacker who thereby successfully answers security questions, often through informed guesses, can reset victims' passwords himself.

T9. Disguise as User
In case of Korean portals, user authentication at password reset- step 2 is conducted with the use of resident registration no. by receiving either Resident Registration no. or copy of identification card. However, frequent spill of personal data including Resident Registration no. makes us doubt about the effectiveness of system to conduct user authentication based on consistency of Resident Registration no. and name [17]. The attacker who obtained the Resident Registration no. of attack target can reset password after sending personal data via e-mail or fax by disguising as the attack target.

4.2 Security Requirements for Authentication System of Portals and SNS

Security affecting the security vulnerabilities of portals' and SNSs' password authentication systems at each stage, as discussed previously, are shown in Table 6.
R1. CAPTCHA

CAPTCHA is a method used to distinguish whether the user is an actual person or a computer program, using something easily distinguished by people but not computers, such as the contents of a picture showing intentionally distorted or overlapping letters [16]. Unmanned registration or authentication programs are executed automatically by computers rather than people so these automated attacks, which may try to create or access accounts, are blocked by CAPTCHAs. A complete enumeration survey attack is an attack that obtains the correct password through random substitution of the password mainly with the use of a computer program. In order to acquire portal

accounts, the attacker can execute a complete enumeration survey program for consecutive login attempts. Therefore, the attack using complete enumeration survey program cannot be blocked in case of requesting the input of CAPTCHA at login.

Table 6. Security requirements that accommodate security threats

the Phasing of Security Threat / Security Requirement	Member	Login			Password Reset – Phase 1			Password Reset – Phase 2	
	T1.Automatic Registration	T2. Consecutive Login Attempts	T3. Phishing	T4. Keylogging	T5. Consecutive Login Attempts	T6.Email Sniffing	T7.Eavesdrop Smart Phone	T8.User Information Guessing	T9. Disguise as user
R1. CAPTCHA	×	×			×				
R2. Password with Enhanced Security Strength		×			×				
R3. Two channel authentication		×		×	×				
R4. Anti-Keyboard Hacking Program				×					
R5. Virtual Keyboard				×					
R6. Login IP Address Identification		×	×		×				
R7. Overseas IP Address Block		×	×						
R8. Anti-phishing and Countermeasures			×		×				
R9. Account Lock		×							
R10. Encrypted communication			×		×	×			
R11. Strength of Security Questions for Password Reset								×	×
R12. Installation of Vaccine Program (User)							×		

R2. Password with Enhanced Security Strength

The time it takes to crack a password and the difficulty of a complete enumeration survey attack is related to the user's password strength. For a user to create a password secure enough for a complete enumeration survey attack, the password should satisfy the following conditions [19].

- The inclusion of both upper and lower case letters, numbers, and special characters
- A minimum of 8 characters
- The prohibition of passwords based on guessable personal data such as the names of family members, phone numbers, etc.
- The prohibition of passwords which are the same as for other web sites

R3. Two channel authentication

Two channel authentication improves on the weak security of single channel authentication using a combination of two different authentication channels chosen from three sources: information possessed by the user, unique information or known information. The most common method is to

combine knowable information such as a password with possessed information such as OTPs, security tokens or smart phones. This approach can avoid the damage caused by ID theft through remote access.

R4. Anti-Keyboard Hacking Program

Keylogging refers to intercepting and recording the contents of users' input on either PCs or smart phones and its various methods may be based either in hardware or software, and include electronic or even acoustic technology [20]. Keylogging programs, hereafter referred to as keyloggers, are difficult to detect and delete once installed so users should take care not to install malicious programs. Vaccine programs and anti-keyboard hacking programs block the attacker from obtaining the ID and password of the user based on keyboard input. One method is to install a special security keyboard driver which outputs special characters, including '*' amongst others, to a security input window connected to the keyboard security driver and thereby transmits null values into the previous keyboard input stream so that no meaningful keyboard input can be intercepted. The second method is for a user to transmit an encrypted value from a separately installed keyboard security driver every time the user enters values into an input window with a new encryption key being created each time the user selects an input window. In this case, even if the attacker obtains the keyboard input values, he cannot know which value is associated with which true keyboard value as the stream is encrypted. The last method of evading keyloggers is instructing users to click input values with a mouse in a virtual keyboard window on the PC screen in case a keylogger is currently running. By installing these software-based technologies, users can block keyboard hacking programs.

R5. Virtual Keyboard

A virtual keyboard is a keyboard presented on screen for the input of passwords for public key certificates, and account passwords, amongst others, and is mainly used for banking transactions. Users enter input values through the on screen keyboard with a mouse click or touch in the case of smart phones or tablets. Since the keyboard structure of a virtual keyboard is created randomly, the actual value entered is not exposed even when the coordinate values are known. Thus, the attacker cannot easily obtain the actual value input from the encrypted format transmitted. Thus, password exposure can be prevented with this method even if the attacker attempts to obtain user passwords by installing a keylogger.

R6. Login IP Address Identification

User authentication is requested if the login IP deviates from the range of IP addresses saved from previous logins or if the IP addresses are different from the one of the last login. Users who succeed in user authentication are recognized as normal users and allowed account access while others are considered as attackers and denied account access.

R7. Overseas IP Address Block

SPAM mail is mostly sent from China and the account hacking of normal users in a given region is normally done from servers in that same region [21]. Therefore, portals should respond to related possible attack attempts by allowing the access of normal users through user authentication stages and then informing them of overseas login attempts from countries that they have not registered.

R8. Anti-phishing and Countermeasures

Anti-phishing methods include blocking sites presumed to be used for phishing after their detection and training users to distinguish phishing sites from normal ones. Phishing site detection methods are largely divided into searches for similar domains and HTTP traffic analysis. Phishing site detection through domain similarity can be classified into blacklist and whitelist techniques. Blacklist-based detection techniques register the addresses of servers known

to be phishing sites and do not trust those addresses included. Whitelist-based detection techniques, on the other hand, register the addresses of legal servers and trust those included. HTTP traffic analysis detects sites which are disguised using links of pictures and postings from normal ones by monitoring and analyzing the HTTP traffic corresponding to requests for postings and pictures from normal sites referred to by the phishing ones.

R9. Account Lock
When an attacker conducts consecutive login attempts using a complete enumeration survey, user accounts will be obtained eventually if there is no restriction on the number of authentication attempts. Therefore, the acquisition of user accounts can be prevented by limiting the number of authentication attempts.

R10. Encrypted Communication
Encrypted communication refers to the transmission of content encrypted by means of a shared key in order to block the tapping or interception of unencrypted content by third parties. Since users who do not possess the key cannot access the plaintext, data spill can be prevented even when the packets themselves are exposed. Therefore, portals should provide encrypted communication for confidentiality, integrity, and user authentication of communications between entities.

R11. Strength of Security Questions for Password Reset
The types of security questions provided in the past were either easy for attackers to guess so insecure, such as "What is the name of your mother?" or difficult to remember so inconvenient, such as "What is your dream job?" [22]. Thus, service providers should improve user authentication security questions. Additionally, multiple security questions should be asked rather than just one, and a real person should be distinguished from an attacker by the percentage of correct answers given. Also, to improve user convenience the questions should be based on their experiences and behaviors when using their accounts so that, rather than having to actively remember the answers, a real user would just know them as a matter of course.

R12. Installation of Vaccine Program (User)
Secure smart phone use is possible when users install vaccine programs on smart phones which conduct regular inspection to pre-empt problems such as mobile phone tapping and data leakage, amongst others, which are caused by attackers using malicious programs.

5. MEASUREMENT OF THE POSSIBILITY OF SUCCESSFUL ATTACKS BY PORTAL SITE AND COMPARISON OF THE SAFETY

This chapter measures the possibility of successful attacks on the steps of the password authentication systems of each portal site based on the possibility indicators of successful attacks based on the common criteria mentioned in 2.2 and compares and analyzes their safety.

5.1. Measurement of the possibility of successful attacks by portal and SNS

This section measures the possibility of successful attacks on each portal site based on attack threats on the portal sites' password authentication systems and the security requirements created earlier.

5.1.1. Member registration

At the member registration stage, account creation using automatic member registration programs is the main attack threat. At the member registration stage, attack scenarios are similar across all portal sites because their security threats and countermeasures are similar. It takes less than one

hour to create an account using automatic member registration programs. Furthermore, the general public can operate such programs because they do not require a deep knowledge of security. To prevent such attacks CAPTCHA and cell phone authentication is used. This policy is the information which was already opened. Automatic member registration programs which can overcome CAPTCHAs are classified as 'professional equipment' because they are available to only a small number of people in specific internet communities.

There are differences between Korean and overseas portal sites in terms of vulnerability to attack. It is impossible to continually create accounts on Korean portal sites, even when using automatic member registration programs, because users are restricted to three accounts per cell phone. However, overseas portal site registrations lack this restriction.

Table 7. Attack Potential of Membership Registration

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	1	1	1	1	1	1	1	1
Expertise	0	0	0	0	0	0	0	0
Knowledge of Object	0	0	0	0	0	0	0	0
Access to Object	1	1	1	0	0	0	0	0
Tools	7	7	7	7	7	7	7	7
Total	**9**	**9**	**9**	**8**	**8**	**8**	**8**	**8**

5.1.2. Log-in

Possible attack methods at the log-in stage include consecutive authentication attempts, phishing, and key logging. Attack methods are classified by attack technique when measuring the possibility of successful attacks. Since a system's overall vulnerability to attack is based on its weakest point, the overall possibility of a successful attack at each log-in stage is based on the attack technique with the minimum score.

5.1.2.1. Consecutive authentication attempts

The total time required for consecutive authentication attempts to be successful are calculated based on each portal site's password strength:

Table 8. Elapsed time for brute force attack

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	13 mins	1 hour 32 mins	6 days 4 hours	6 days 4 hours	82 days 21 hours	17 years 130 days	13 mins	13 mins

In Table 10 above, the time measurement was calculated supposing that the 'John the Ripper', attack tool, is run on an attacker's PC with a 3.4GHz Intel Core i7-2600K and assumes the use of the simplest password allowed. Naver, Facebook and Twitter take less time to attack because they use only 6-digit passwords and do not provide an account lock service. Yahoo takes the longest time because it makes use of a compulsory combination of upper and lower case letters as well as numbers. Naver and Nate were easier targets than other portal sites because they do not provide an account lock service.

Attack tools for consecutive authentication attempts are shown in Table 9. Everybody can obtain them because they are easily available on the internet and experts with enough knowledge about security can use them.

Table 9. Attack Potential of Consecutive Authentication Attempts

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	1	3	5	5	10	15	1	1
Expertise	6	6	6	6	6	6	6	6
Knowledge of Object	0	0	0	0	0	0	0	0
Access to Object	1	1	4	4	4	4	4	4
Tools	4	4	4	4	4	4	4	4
Total	**12**	**14**	**19**	**19**	**24**	**29**	**15**	**15**

5.1.2.2. Phishing

Phishing's probability of success depends on the similarity between the original sites and the special phishing sites built by attackers. If the source code of the log-in screen of a portal site is exposed, anybody can build the phishing site simply by copying the code. Otherwise, the phishing sites need to be built with web site building tools so as to be as similar as possible to the original. This process takes more time but usually only requires basic web knowledge. Google, MSN, Facebook and Twitter, take longer to attack than other portals because the source code of their log-in pages is not exposed. Because sites built by coping page sources are more similar to the original sites, attackers can obtain the passwords of more targets. Naver, Google, and MSN are applying anti-phishing technologies which give warnings about phishing or malware sites to not only tool bars, such as MSN Tool Bar-phishing filter and Naver anti phishing Toolbar, but also to browsers, such as Google Chrome. In addition, Yahoo provides a security seal service which is a phishing prevention technology that allows users to recognize that they are accessing the real site because pictures chosen by themselves in advance are presented at the log-in stage. Therefore, attacks on Naver, Google, MSN, or Yahoo are more difficult to create than those on Nate or Daum, when users take advantage of their anti-phishing technologies.

Table 10. Attack Potential of Phishing

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	3	3	3	5	5	3	5	5
Expertise	3	3	3	6	6	3	3	3
Knowledge of Object	0	0	0	3	3	0	3	3
Access to Object	4	1	1	4	4	4	1	1
Tools	4	4	4	4	4	4	4	4
Total	**14**	**11**	**11**	**22**	**22**	**17**	**16**	**16**

5.1.2.3. Key logging

All the portal sites have the same likelihood of successful attacks by key logging because it is controlled by users' browsing environments rather than a portal's security policies. User accounts can be obtained if users allow key logger programs access to run on their PCs and harvest and interpret ID and password values. The time required for key logging attacks depends on time needed to interpret the key values entered, and would usually be less than one day. Key logger programs are openly available on the internet and using them attackers can easily access the accounts of targets who read the malicious texts or spam mails which disseminate keyloggers.

Table 11. Attack Potential of Keystroke Logging

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	3	3	3	3	3	3	3	3
Expertise	3	3	3	3	3	3	3	3
Knowledge of Object	0	0	0	0	0	0	0	0
Access to Object	1	1	1	1	1	1	1	1
Tools	4	4	4	4	4	4	4	4
Total	**11**	**11**	**11**	**11**	**11**	**11**	**11**	**11**

The result showed that key logging's possibility of successful attacks is the lowest. Furthermore, all portal sites have the same vulnerability to this type of attack at the log-in stage.

5.1.3. Password reset - Phase 1

If users request a password reset, attackers can obtain their accounts by three methods, SMS wiretapping, complete enumeration surveys, and access via other e-mail accounts. Therefore, the possibility of successful attacks in the password reset step 1 of each portal site is based on the attack technique with the minimum score since security can only be as strong as its point of weakest defence.

5.1.3.1. SMS wiretapping

SMS wiretapping obtains authentication numbers as they are delivered to users' cell phones by installing wiretapping applications when users inadvertently install them during regular cell phone use. This method is effective when attacks are specifically targeted.

Malicious wiretapping applications can be created or purchased by attackers and ordinary people can easily use them. Because the authentication numbers transmitted to users by SMS for authentication via cell phone are valid for three minutes, attacks can be successful only if the authentication numbers can be obtained and the passwords reset within this time limit. Information about cell phones' weak points is considered to be openly available information because it can be obtained through on-line searches. As the success of these attacks is determined by the functionality of the malicious applications installed on users' cell phones, portals' vulnerabilities to such attacks are unaffected by their security policies. Thus, for all portal sites, the probability of this attack type being successful is the same.

Table 12. Attack Potential SMS Eavesdropping attacks

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	1	1	1	1	1	1	1	1
Expertise	0	0	0	0	0	0	0	0
Knowledge of Object	0	0	0	0	0	0	0	0
Access to Object	4	4	4	4	4	4	4	4
Tools	4	4	4	4	4	4	4	4
Total	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**

5.1.3.2. Complete enumeration surveys of authentication numbers

It takes less time to conduct complete enumeration surveys as the number of cases is less because authentication numbers consist of 6-digit numbers. Portal sites provide account lock services if users fail in consecutive authentication of authentication numbers to respond to the complete

enumeration surveys. The detailed contents are shown in Table 13. However, the accessibility to authentication numbers is difficult because there are the only five to ten chances to enter them in one million 6-digit numbers of cases as they are randomly transmitted in request repeat unlike the fixed passwords. And it can be found that it takes more than 6 months to attack authentication numbers.

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Input	5 times	5 times	5 times	3 times	3 times	3 times	3 times	3 times
Transmission	10 times	10 times	10 times	5 times	5 times	5 times	5 times	5 times
Account Lock	24 hours	24 hours	24 hours	24 hours	24 hours	24 hours	24 hours	24 hours

Although the tools, knowledge levels, and skill required for complete enumeration survey attacks on authentication numbers are the same as for consecutive authentication attempts, the time limit and targets of attacks depend on the policies of particular portal sites regarding the authentication number and its input or transmission.

Table 13. Attack Potential of Bruteforce Validation Code

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	15	15	15	15	15	15	15	15
Expertise	3	3	3	3	3	3	3	3
Knowledge of Object	0	0	0	0	0	0	0	0
Access to Object	10	10	10	10	10	10	10	10
Tools	4	4	4	4	4	4	4	4
Total	**32**	**32**	**32**	**32**	**32**	**32**	**32**	**32**

5.1.3.3. Accessibility to other e-mail accounts

Access via other e-mail accounts is a possible attack method if users choose to identify themselves for the purpose of password resets through e-mail authentication. Each portal site uses e-mail authentication methods like Table 17. Yahoo is immune to this attack because it only accepts cell phone authentication so email authentication is impossible.

Table 14. E-mail Authentication

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Authentication Method	Code	URL	Code	URL	Code	-	URL	URL
Open	1st character	All	1st character	1st character	All	-	1st character	2 characters

The time required for attacks through access via other e-mail accounts depends on portal sites' distribution of information to other accounts. Log-in attempts on the accounts used for this type of attack have the same possibility of success as normal log-in attacks on these sites. Because Nate and MSN make available the information of other accounts which transmit emails to them for identification, and it takes a short time to attack these portals, accessibility to other accounts is easier through them as attackers can attempt attacks on related accounts without assumption. Furthermore, for attacks via other accounts the attackers can use information they have obtained by themselves directly rather than though specific tools, so specialized security knowledge is not required.

Table 15. Attack Potential of E-mail Account attack

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	3	1	3	3	1	-	3	3
Expertise	0	0	0	0	0	-	0	0
Knowledge of Object	0	0	0	0	0	-	0	0
Access to Object	4	1	4	4	1		4	4
Tools	0	0	0	0	0	-	0	0
Total	**7**	**2**	**7**	**7**	**2**	**-**	**7**	**7**

From the previous analysis of the possibility of successful attacks during password reset Step 1, it can be concluded that the possibility of successful attacks using accessibility from other e-mail accounts is higher than the other methods. Therefore, the possibility of successful attacks in password reset step 1 is the same as the possibility of successful attacks exploiting the accessibility from other e-mail accounts.

5.2.4. Password reset – Step 2

Until August 2013 korean portal sites had provided for identification processes that utilize social security numbers. However, they had not provided phase 2 authentication services. To compare the safety of Korean and overseas portal sites, this paper measured the possibility of successful attacks on identification processes through the collection of social security numbers or information to answer social security questions.

Social security numbers are available to only a small number of people through distributors. The attackers who have obtained the social security numbers of attack targets can obtain their account information and receive their passwords from portal sites by forging or using them. It takes less than one week to attack targets, including searching for their IDs and checking their social security numbers and, furthermore, this attack can be easily carried out without specialist security knowledge.

The overseas portal sites verify users through questions based on their experiences or behavior during account use. Although attackers can guess the answers after collating information about users readily available on the internet, it takes experts about one week to collect and analyze the information. Special tools for the attacks are not required. However, the answers to the questions are considered important information because it is information which is remembered and known without special effort.

Table 16. Attack Potential of Password Reset-2nd Phase

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Elapsed Time	5	5	5	5	5	5	5	5
Expertise	0	0	0	3	3	3	3	3
Knowledge of Object	3	3	3	11	11	11	11	11
Access to Object	4	4	4	4	4	4	4	4
Tools	0	0	0	0	0	0	0	0
Total	**12**	**12**	**12**	**23**	**23**	**23**	**23**	**23**

5.2. Comparison and analysis of password authentication systems of portal sites based on the possibility of successful attacks

The safety of password authentication systems of portal sites is compared and analyzed based on the possibility of successful attacks measured in the previous section. The table which finally

analyzed the possibility of successful attacks by step is as follow. Comparison of the safety by step of password authentication systems was analyzed that the safety of password reset Step 1 is lowest. It's because the vulnerability in the log-in stage is frequently used in attacks while attackers can more easily obtain passwords when it is actually done.

Table 17 Attack Score of Password Authentication Systems

	Naver	Nate	Daum	Google	MSN	Yahoo	Facebook	Twitter
Membership Registration	9	9	9	8	8	8	8	8
Login	11	11	11	11	11	11	11	11
Authentication phase 1	7	2	7	7	2	9	7	7
Authentication phase 2	12	12	12	23	23	23	23	23

A method to compare the safety of the Korean and overseas portal sites by step and strengthen them is as follow. The member registration stage can be found that the Korean portal sites are safer than the overseas ones because they limit the number of IDs. The possibility of successful attacks about attack scenarios of consecutive authentication attempts, key logging, and phishing was measured in the log-in stage. Key logging was analyzed to be all the same values in all the portal sites. The Korean portal sites and two of SNSs have found to be vulnerable to key logging attacks because they do not provide the methods which can respond to key logging attacks. The overseas portal sites can be found to be more vulnerable to consecutive authentication attempts than the Korean ones. It's because Facebook, Twitter, Naver and Nate do not provide account lock services and their password strength is lower than that of the overseas portal sites. They should improve their password strength and provide the account lock services to complement this. Phishing attacks of Naver, Google, MSN, and Yahoo have found to be safer than Nate, Daum, Facebook and Twitter because they provide the technologies to respond to them. However, the Korean portal sites and Yahoo were analyzed that phishing sites can easily be built because their source codes are exposed. They need to make their source codes' interpretation difficult to complement them and Naver and Daum should provide their own anti-phishing technologies.

Because possible SMS wiretapping attacks in password resent step 1 are related to the safety of user smart phones, users can be safe from the applicable attacks as it is recommended to install vaccine in them. Complete enumeration attacks of authentication numbers could be found to be safe because the Korean and overseas portal sites all limit the number of input and transmission of authentication numbers. If specific users are targets of attacks, attackers can more easily obtain passwords than the log-in stage by utilizing the accessibility to them through other e-mail accounts of the authentication stage through the e-mails in password reset step 1. If other accounts registered by users are exposed, the time required can be more reduced than the cases that they are not opened. And if password strength of the registered accounts is lower than that of the accounts that users try to reissue, the attack level of difficulty get to be lower if attackers attack the applicable accounts. Therefore, Nate and MSN that other account information is exposed should improve the safety just by exposing the partial accounts.

For the password reset step 2 service of the Korean portal sites, the authentication method which use social security numbers before Aug. 2014 was analyzed. As personal information leaks frequently occurred, the social security numbers are circulating the market. And attackers can easily obtain them. Therefore, password reset by attackers disguised as users had no major difficulties. The overseas portal sites are safer than the Korean ones as it is difficult for attackers to guess right answers because the information contents of users are not opened because they are based on the experiences that they just know.

6. CONCLUSION

This paper analyzed the authentication systems of Naver, Nate, Daum, Google, MSN, and Yahoo, the main Korean and overseas portal sites, and clarified existing security threats and the security improvements necessary to remedy them, and analyzed the application of the security requirements which were drawn in each portal site to them. From the analysis of the password authentication systems of the Korean and overseas portal sites it was found that the Korean portals Naver and Nate are more vulnerable to complete enumeration attacks than the overseas ones because they do not provide account lock services. However, it was also found that the foreign portals did not provide a service blocking logins or the identification services from overseas, except Google. This may be because for users with malicious intent, the creation of new accounts is preferred over attempts to seize existing users' accounts because the number of multiple IDs which can be created has no limit. However, because the motivation for attacks seizing users' accounts is not only their acquisition for sending spam mails but also accessing the private user information they contain, protection against this should be improved. Furthermore it has been found that both Korean and overseas portal sites do not force users to make safe passwords. Because exploiting weak passwords is an attack method not only at log-in but also password reset step 1, all the portal sites should make creating safe passwords compulsory. The Korean portals should also improve the convenience of users who do not use their cell phones for authentication at password reset step 2 by using user behavior-based authentication methods or develop new methods as the overseas portal sites have done.

This paper analyzed the entire process of password authentication on Korean and overseas portal sites, explained their potential security vulnerabilities, and proposed security-hardening solutions for each process. Moreover, it quantitatively compared and analyzed the safety of the password authentication systems of the major Korean and overseas portal sites by the creation and use of a standardized set of criteria to express the possibility of successful attacks.

ACKNOWLEDGEMENTS

This work was supported by the ICT R&D program of MSIP/IITP. [2014(10043959), Development of EAL4 level military fusion security solution for protecting against unauthorized accesses and ensuring a trusted execution environment in mobile devices]

REFERENCES

[1] Bruce Schneier, "Applied Cryptography", John Wiley & Sons, 1996.
[2] Perlman, Radia, and Charlie Kaufman. "User-centric PKI". Proceedings of the 7th Symposium on Identity and Trust on the Internet. ACM, 2008.
[3] Just, Mike, and David Aspinall. "Personal Choice and Challenge Questions: A Security and Usability Assessment". Poceedings of the 5th Symposium on Usable Privacy and Security. ACM, 2009.
[4] Jin, Lei, Hassan Takabi, and James BD Joshi. "Analysing security and privacy issues of using e-mail address as identity." International Journal of Information Privacy, Security and Integrity, 1.1. 34-58. 2011.
[5] C.E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, 1948, pp. 379–423.
[6] Komanduri, Saranga, et al. "Of passwords and people: measuring the effect of password-composition policies." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011.
[7] Ma, Wanli, et al. "Password entropy and password quality." Network and System Security (NSS), 2010 4th International Conference on. IEEE, 2010.
[8] Yan, Jianxin Jeff. "A note on proactive password checking." Proceedings of the 2001 workshop on New security paradigms. ACM, 2001.

[9] Bishop, Matt. "Proactive password checking." 4th Workshop on Computer Security Incident Handling. 1992.
[10] “Common Methodology for Information Technology Security Evaluation.” Common Criteria, Version 3.1. 2009.07.
[11] Cazier, Joseph A., and B. Dawn Medlin. "Password security: An empirical investigation into e-commerce passwords and their crack times." Information Systems Security 15.6. pp.45-55. 2006.
[12] Ji Sun Shin, “Study on Anti-Phishing Solutions, Related Researches and Future Directions,” Journal of The Korea Institute of Information Security & Cryptology, Vol.23, No.6, Dec.2013.
[13] Leijten, Mariëlle, and Luuk Van Waes. "Keystroke Logging in Writing Research Using Inputlog to Analyze and Visualize Writing Processes." Written Communication 30.3, pp.358-392, 2013.
[14] “2014 Trustwave Global Security Report“, Trustwave, 2014
[15] Dell'Amico, Matteo, Pietro Michiardi, and Yves Roudier. “Password strength: An empirical analysis.” INFOCOM, 2010 Proceedings IEEE. IEEE, 2010.
[16] Irani, Danesh, et al. "Modeling unintended personal-information leakage from multiple online social networks." Internet Computing, IEEE 15.3 ,pp.13-19. 2011.
[17] HyeongKyu Lee, “The Problems and Reformation of the Personal Identification by the Resident Registration Number on the Internet”, Hanyang Law Review, Vol. 23-1, pp.341~371, 2012. February.
[18] Von Ahn, Luis, et al. “CAPTCHA: Using hard AI problems for security.” Advances in Cryptology—EUROCRYPT 2003. Springer Berlin Heidelberg, 2003. 294-311.
[19] Jin, Lei and Takabi, Hassan and Joshi, James B.D, “Analysing security and privacy issues of using e-mail address as identity,” International Journal of Information Privacy, Security and Integrity, 1 (1). pp. 34-58. 2011.
[20] Goring, Stuart P., Joseph R. Rabaiotti, and Antonia J. Jones. "Anti-keylogging measures for secure Internet login: an example of the law of unintended consequences." Computers & Security 26.6. pp.421-426. 2007.
[21] “Kaspersky Releases Q1 Spam Report,” Kaspersky, 2014.
[22] Jin, Lei and Takabi, Hassan and Joshi, James B.D, “Analysing security and privacy issues of using e-mail address as identity,” International Journal of Information Privacy, Security and Integrity, 1 (1). pp. 34-58. 2011

AUTHORS

Heekyeong Noh received her B.S degree in Internet Information Engineering from Duksung Women’s Unviersity of Korea, in 2012. She is currently working toward M.S degree in Information Security, Korea University(KU), Korea. Her research interests include password security, security engineering, and Common Criteria(CC)

Changkuk Choi received his B.S degree in Department of Chemical Engineering from Kwang Woon University of Korea, in 2000. He is currently working toward Ph.D degree in Information Security, Korea University(KU), Korea. His research interests include hacking, CCTV security.

Minsu Park received his B.S degree in Computer Network from Silla University of Korea, in 2010 and also received his M.S degree in Information Security from Korea University(KU) of Korea in 2013. He is currently working toward Ph.D degree in Information Security, Korea University(KU), Korea. His research interests include information assuarance, digital forensic, and usable security.

Jaeki Kim received his B.S. (2013) in Computer Engineering from Hanyang University ERICA in Korea. and, He served as Security Technology Team of the INetCop for 1 years. also, He participated a program for the training next-generation's best IT security leaders, called 'Best of the Best' 2nd (2013). His research interests include Android Security and Embedded devices Security. He is now a graduate student at CIST SANE LAB, Korea University.

Seungjoo Kim received his B.S., M.S. and Ph.D. from Sungkyunkwan University (SKKU) of Korea, in 1994, 1996 and 1999, respectively. Prior to joining the faculty at Korea University (KU) in 2011, He served as Assistant & Associate Professor at SKKU for 7 years. Before that, He served as Director of the Cryptographic Technology Team and the (CC-based) IT Security Evaluation Team of the Korea Internet & Security Agency (KISA) for 5 years. He is currently a Professor in the Graduate School of Information Security at KU, and a member of KU's Center for Information Security Technologies (CIST). Also, He is a Founder and Advisory Director of a hacker group, HARU and an international security & hacking conference, SECUINSIDE. Prof. Seungjoo Kim's research interests are mainly on cryptography, Cyber-Physical Security, IoT Security, and HCI Security. He is a corresponding author.

STUDY ON ANALYSIS OF COMMERCIAL MOBILE KEYPAD SCHEMES AND MODELING OF SHOULDER SURFING ATTACK

Sunghwan Kim[1], Heekyeong Noh[2], Chunghan Kim[3], and Seungjoo Kim[4]

CIST (Center for Information Security Technologies),
Korea University, Seoul, Korea
[1]tonykimsh@korea.ac.kr, [2]hknoh@korea.ac.kr,
[3]6sasimi@korea.ac.kr, [4]skim71@korea.ac.kr

ABSTRACT

As the use of smart phones and tablet PCs has exploded in recent years, there are many occasions where such devices are used for treating sensitive data such as financial transactions. Naturally, many types of attacks have evolved that target these devices. An attacker can capture a password by direct observation without using any skills in cracking. This is referred to as shoulder surfing and is one of the most effective methods. There is currently only a crude definition of shoulder surfing. For example, the Common Evaluation Methodology (CEM) attack potential of Common Criteria (CC), an international standard, does not quantitatively express the strength of an authentication method against shoulder surfing. In this paper, we introduce a shoulder surfing risk calculation method that supplements CC. Risk is calculated first by checking vulnerability conditions one by one and the method of the CC attack potential is applied for quantitative expression. We present a case study for security-enhanced qwerty-keypad and numeric-keypad input methods, and the commercially used mobile banking applications are analyzed for shoulder surfing risks.

KEYWORDS

Shoulder surfing attack, Attack potential, Security keypad

1. INTRODUCTION

As technologies such as the Internet, mobile, and wireless communications developed internationally, mobile devices became gradually smaller, and communications became faster. Accordingly, mobile devices evolved into the era of the smart phone, and users came to be able to develop various applications using open SDK (Software Development Kit) and then to use various contents. Referring to BOK (Bank of Korea) data, with the continued increase in the use of smart phones since the end of 2009, there were 37 million subscribers in South Korea by January 2014, and this growth has continued [1]. In addition to smart phones, the use of tablet PCs, too, has consistently increased, and there were about 650,000 subscribers as of January 2014 [2]. As a result, as smart phones and tablet PCs with mobility, portability, and convenience become distributed smoothly and there is much usage, more people enter or process important information using the relevant smart devices. Accordingly, virus infection through spyware created in the PC environment, malware installation through illegal file downloads, MITM (man in the middle attack) that intercepts user information, keystroke logging attack, and attack using social engineering techniques have all moved to smart devices. These attack techniques mostly aim at financial applications that may cause an economic loss, and attacks mostly target a user's important password such as a certificate password or account transfer password. The number and

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 93–112, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41208

value of uses of mobile banking in the first quarter of 2014 were 27.6 million and 1.6634 trillion won, respectively, which were increases of nine thousand and about four hundred billion won respectively from the first quarter of 2013, and because it appears that usage will continue to increase consistently, careful attention among users is necessary [1]. In addition, in spite of an attempt to express shoulder surfing attack quantitatively by applying attack potential, it is erroneous to judge or measure the rating of attack potential in detail based on the elements of attack potential of existing CEM (Common Evaluation Methodology). This is because existing methodology does not include attack elements reflecting the characteristics of shoulder surfing attack. So there currently exists no standard by which the tolerance to shoulder surfing attack on the password input scheme can be judged.

Thus, this study proposes an attack potential that adds the attack elements through which attack potential of shoulder surfing attack can be judged to existing elements of attack potential on the password input scheme. In addition, as cases of the application of the proposed attack potential, with the password input scheme of the mobile banking applications provided in the market as subjects of analysis, this study analyzes the status of safety to determine whether the security keypads of mobile banking applications are safe from shoulder surfing attack.

2. RELATED WORK

2.1. Shoulder Surfing Attack and Password Input Scheme

Shoulder surfing attack refers to peeping around a user who is logging in or looking at sensitive information, without their awareness, when the user uses specific devices (smart phones, laptops, or PDAs) at an office, crowded shopping mall, airport, or coffee shop [3]. Thus, shoulder surfing attack is a powerful and effective means of attack for clearly observing the user password. In response to this, studies have been carried out in order either to develop an input scheme that allows users to enter their password behind their smart phones so that they are safer from shoulder surfing attack than with the existing password entry methods or to design and implement a method using a CoverPad to prevent information leak by attacks such as peeping when the users enter their password by the touch screen method [27][31]. In particular, the study that after standardized modeling of the shoulder surfing attack that is difficult to express by standardization, using a method like CPM-GOMS model, which can express it quantitatively and reviewed and tested the usability and safety of the qwerty keypad is one of the standardized studies related to shoulder surfing attack [26]. Xiaoyuan Suo et al. proposed that a password that a Web or smart phone user should enter for user authentication can be divided into a text-based one and a picture-based one [4]. The text-based password refers to an alphanumeric one consisting of numbers and characters while a picture-based password is divided into a perception-based one in which a user selects or passes one of the sets of passwords selected and registered in the procedure of password registration and a memory-based one in which a user is asked to copy or reproduce the picture created or selected by the user from among sets of pictures following the procedure of password registration. Both types of passwords aim to generate a password that the user can memorize easily and that has high security. While password generation in terms of security and usefulness is important, existing studies that analyzed the qwerty keypads used by current mobile banking applications report that the current qwerty keypads are exposed to the problem of the possibility of abuse by keystroke logging drawing random layouts through stochastic analysis [18], and thus the password security is under threat. Most studies suggest as alternatives to this to propose the use of a type of picture-based password as a substitute for the current security keypad [24]. Focusing on human memory and security, the picture-based password aims to increase the memorability of the password and, at the same time, increase security. Picture-based passwords can be concretely divided into three types: memory-based, perception-based, and memory-based with a clue. These picture-based passwords satisfy both

usefulness and security and are intended for a design safe from several password attacks [25]. However, even if one uses the picture-based password, because of the characteristics of picture-based passwords, there are both safe and unsafe passwords in the case of shoulder surfing attack. First, Sobrade and Birget and Man et al., who proposed a graphical password as one of the types of perception-based picture passwords, argued that the graphical password would be safe from shoulder surfing attack since it is more difficult for an attacker to be able to recognize it as compared with existing text-based passwords [5][6], and Real User Corporation proposed that the Passface password, in which a user registers pictures of four faces in advance and chooses those four pictures from among nine pictures of faces for user authentication, would be safer than existing text-based passwords from shoulder surfing attack [7]. However, the above picture-based password has suitability problems when it comes to the essential characteristics of a password, which must be easy for the user to remember: e.g., one deficiency is that the user has to remember the character string value given to the image or has to memorize the pictures of the four pre-registered faces. Secondly, in a memory-based picture password, Jermyn et al. proposed the Draw-A-Secret (DAS) method of authentication in which a user draws simple pictures on a 2D grid, saves them in order, and then draws them in the same order for successful authentication [8]. And Goldberg et al. proposed the Passdoodle technique, in which a user draws pictures or writes characters randomly on a touch screen [9], and Blonder designed a scheme in which a password is generated by a user clicking on several positions of an image and proposed the Passlogix password technique of authentication by clicking on several positions in the same way to receive authentication [10]. All three password techniques were designed to possess greater resistance capacity against shoulder surfing attack than existing text-based passwords, but DAS, when used on a device with a large screen, may be vulnerable to peeping just as with existing shoulder surfing attack, and Passdoodle and Passlogix are not suitable for easy recall of password, as was the problem with perception-based picture passwords. So these might all lead to problems with users not being able to remember their password. In addition, both types of picture-based passwords were designed to be safe from shoulder surfing attack, but in user feedback, when checking whether the password a real user entered by drawing a picture was entered correctly, the process for validating the password entered before coding or the process of drawing a picture slowly because of unfamiliarity with password input may be exposed to shoulder surfing attack, and this has a weak point similar to that of existing text-based password schemes. Table 1 can be expressed as the schemes described above are features of the password entry.

Consequently, studies to increase security of password schemes, to design safe passwords, and to facilitate a user in remembering the password have consistently been carried out, but developing a password scheme to maintain the balance between safety and usability has been lacking, and there is the problem of vulnerability to shoulder surfing attack. Also, most studies of password schemes until now have depended simply on the password itself without mentioning detailed attack elements regarding shoulder surfing attack. To supplement and improve these points, criteria for judging the status of safety from shoulder surfing attack of password input schemes that have not been presented in existing studies are necessary, and for this purpose, the methodology of attack potential to be introduced in the following will be used.

2.2. Attack Potential

To draw safety criteria for the password input scheme from shoulder surfing attack, this section will follow the methodology of attack potential in the CC (Common Criteria) that previously presented standards so that the status of attack on specific targets such as smart cards or H/W devices can be quantitatively recognized. Attack potential refers to a function of expertise, resource, and motivation presented by CEM in the CC and consists of elapsed time, expertise, knowledge about a target of attack, period of easy exposure to attack, and equipment, and

quantitatively demonstrates the attack potential of the target of attack by assigning values to each element [23].

Table 1. Password input scheme and possible attacks [11]

Password input scheme		Input Method	User Usability	Possible Attacks
Text-based password	Alphanumeric password	Input text and number using keyboard	Problem that making a password easy to remember reduces security	Brute force attack, Dictionary attack Guessing, Malicious program, Shoulder surfing attack
Picture-based password	Graphical password	Click a specific position of the picture registered in advance or enter a specific code passing a few pictures	Problem that it is difficult to remember when there are many other pictures presented together	Brute force attack, Guessing, Malicious program
	Passface	Register 4 pictures of face and select in authentication	Problem that pictures like face with characteristics are easy to remember, but they are predictable	Dictionary attack, Brute force attack, Guessing, Shoulder surfing attack
	Draw-A-Secret (DAS)	Draw and register simple pictures on a 2D grid, and draw them again in order in authentication	Problem that it is difficult for the user to remember the order of drawing	Guessing, Dictionary attack, Shoulder surfing attack
	Passdoodle	Draw a picture randomly using a stylus on the touch screen or enter a text	Easy or difficult to remember depending on what pictures the user draws	Guessing, Dictionary attack, Shoulder surfing attack
	Passlogix	Touch specific parts of a picture in the assigned order for authentication	Problem that it is difficult to remember perfectly	Guessing, Brute force attack, Shoulder surfing attack

The range of the sum of the calculated attack potential values can be expressed using the sub-ranges "0–20," "20–30," "30–34," "over 34," with higher knowledge about the target of attack being accumulated, and with high attack potential, the attacker has high attack potential. By contrast, lower values mean that the attacker has lower attack potential for the target of attack and has lower attack potential.

Table 2. Attack potential using CEM

Attack Potential			
Elements	Description	Standard	Value
Elapsed time	The sum of time taken for an attacker to detect and develop a weak point that may exist in a target of attack and make an effort required for the attack of the target	Within 1 day	0
		Within 1 week	1
		Within 2 weeks	2
		Within 1 month	4
		Within 2 month	7
		Within 3 month	10
		Within 4 month	13
		Within 5 month	15
		Within 6 month	17
		Over 6 months	19
Expertise	General level knowledge about the type of product or attack method	Layman	0
		Proficient	3
		Expert	6
		Multiple expert	8
Knowledge about target of attack	Detailed specialized knowledge related to the target of attack	Public information	0
		Restricted information	3
		Sensitive information	7
		Critical information	11
Period of easy exposure to attack	Period (chance) related to elapsed time, when an attacker can approach the target of attack	Unnecessary/Unlimited access	0
		Easy access	1
		Moderate access	4
		Difficult access	10
Equipment	An attacker can use equipment to detect or take advantage of vulnerability of the target of attack, which is related to specialized knowledge, so the attacker with high specialized knowledge can use equipment with high attack potential.	Standard equipment	0
		Specialized equipment	4
		Customized equipment	7
		Complex customized equipment	9

However, the attack potential presented in CEM has limitations in that it does not include or meet attack elements necessary to quantify shoulder surfing attack, one of the attack methods for the password input scheme, nor does it judge the status of safety, and so it is not suitable for evaluating shoulder surfing attack. For this reason, existing attack potential possesses two problems: difficulty in judging the status of safety from shoulder surfing attack in a password input scheme and lack of reasonable criteria for calculating the attack potential of shoulder surfing attack.

Table 3. Vulnerability level of attack potential

Range of value	Attack potential
0–20	Low
20–30	Medium
30–34	High
Over 34	Very High

3. SHOULDER SURFING ATTACK MODELING

3.1. Necessity of Formalized Attack Modeling

Various attack techniques on passwords have been presented in various ways. Table 1 shows the methods of attack and user usability possible for the above-mentioned picture-based password and text-based password. Various password-related attack methods have continued to be developed and executed until now. These include a) a keystroke logging attack that obtains coordinate information when the password the user enters is transferred to the server by inducing the user to install a malicious program to detect the password, b) an exhaustive search attack to detect user password by attempting all possible combinations of password, or c) a dictionary attack that can detect the key at considerably high probability when applied to a real situation by creating values that may be user key as a huge dictionary. Of them, shoulder surfing attack is possible with all password input schemes, as can be seen in the details presented in Table 1 [11]. This is because shoulder surfing attack can obtain relevant information by obtaining user information directly through peeping.

However, since shoulder surfing attack is one of the ergonomic aspects such as human perception, cognition, viewing angle, and memory, unlike with the other attack methods mentioned, there has been difficulty in quantitative expression. However, if quantitative expression is made possible through formalized modeling of threats and attack environments of attack conditions of shoulder surfing attack, a value of attack potential can be set for a shoulder surfing attack on a particular password input scheme so as to quantify that, and through the corresponding figures, whether the shoulder surfing attack on the password input scheme is safe can be determined. As seen in the cases of analysis applying attack potential to a specific target such as a smart card, ATM device, POI, or H/W device, the status of the attack potential can be judged by presenting the attack target's attack potential rating as scores [28][29][30]. However, in the above cases, as mentioned, there is a critical point in that it is difficult to present the attack elements in shoulder surfing attack using the elements of existing attack potential, so the following requirements should be drawn.

3.2. Drawing and Analyzing Requirements for Formalized Attack Modeling

This section lists the following attack elements to present standards for showing attack potential of shoulder surfing attack on the password input scheme and suggests standard values. Each standard value presented in each element was estimated, reflecting the characteristics that can be adjusted according to the type of technology and specific environment as described in the CEM document [23]. The attack potential of each element was classified into four values (1, 4, 7, 10), as with the CEM values, and the intervals of 3 points are differential values according to the characteristics described in the below attack elements, which aim to give distinction to the impacts of each element on shoulder surfing attack. This is a result of the use of the characteristics of attack potential that may be estimated differently according to the environment of the assessor who judges the tolerance of shoulder surfing attack on the password input scheme as well as the state of possible abuse.

3.2.1 Perception and recognition

In general, people go through processes of perceiving and recognizing texts (characters and numbers). As these processes are necessary and important elements in peeping over the shoulders at a user's important information, an attacker needs the ability to perceive and recognize the user's entering motions fast and accurately. The processes have a few common characteristics: first, recognition of a certain amount of texts in a given time. An English user can usually read and

recognize sentences at a speed of 360 words per minute on average; that is, 6 words per second, which takes approximately 50 milliseconds [19]. Second, the sensory organs necessary for perception and recognition and the storing capacity for remembering this. One should remember the texts obtained through the processes of perception and recognition for a short time, which is called short-term memory and refers to memory that keeps one's experiences in one's consciousness for several seconds. Miller argued that the memory capacity of humans is 7 ± 2 items [32], and in a later study, he insisted that they memorize things in chunks, for example, they memorize about seven numbers, about six characters, and about five words as a chunk and immediately memorize what they need, dividing this into detail [33]. Since short-term memory carries out storing, perceived information can be said to be an important part of success in shoulder surfing attack, and so the capacity of short-term memory in an attack should be considered sufficiently. Lastly, the items stored in the above-described short-term memory may last for a short time or not be moved to long-term memory simply because of the lapsing of time and be forgotten within a short time. At this time, the recall rate for the items presented first may exhibit a high recency effect, which may act as important elements of how efficiently the attacker recalls the information peeped over the shoulder [12].

Consequently, when the user enters the password through the security keypad, the shoulder surfing attacker may obtain the user password through processes of perception and recognition only by his or her own sensory organs. Thus, the time taken for the attacker's perception and recognition should be shorter than that taken by the user between the input of the first password and that of the last one for a successful shoulder surfing attack. The existing GOMS-based model was used for user interface modeling and assumes perception, recognition, and behavior manipulations and predicts the run time through the critical path. As shown in the STM-GOMS study that applied this to shoulder surfing attack, the time, including the standby time such as that for the movement of an eye and a finger, is approximately 5.8 sec (5,840 msec.) while the total time of the input is approximately 5.3 sec (5,280 msec.). Regarding this, the time taken for the attacker's perception and recognition, including the movement of the eye that looks at the key the user presses, is 180 msec., the total time the attacker waits for the user's input during the total execution time is approximately 2.5 sec (2,510 msec.), and the time of the attacker's actual attack is approximately 2.8 sec (2,820 msec.) [19]. In this way, if the attacker's actual attack time is shorter than the total time of the user's input of password, shoulder surfing attack can be successful, and attack potential may appear, depending on the attacker's time of actual attack.

Table 4. Attack potential about perception and recognition

Perception and Recognition Time	Value
Over 5.8 sec.	1
4.8–5.8 sec.	4
3.8–4.8 sec.	7
2.8–3.8 sec.	10

3.2.2 Screen angle

People generally gaze at the screen of a smart phone reflected by light or from a looming and invisible angle. The angle from which they can see the contents of the screen best is when the screen is located horizontally and forms a 90° angle with the user's gaze—when the user is located as in Fig. 1—and the range of a total of 270°, excluding 315 to 45° hidden by the body, is the angle from which he or she can peep at the information on the screen through shoulder surfing attack [13].

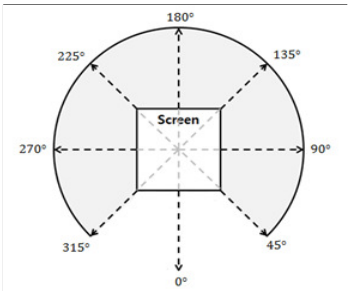


Figure 1. Horizontal angle of screen

A total of 270° can be said to be the angles from which the attacker can peep best. The 360° around the user are divided by 45°, and the angles in bilateral symmetry are evaluated as the same from 180°. In the horizontal angle between 0 and 45° (315 and 360°), from which the user looks at the screen, the screen is hidden by the user, so the attacker cannot see the screen, and thus it is judged that there is no attack potential. In the angle between 45 and 90° (270 and 315°), though oblique, the attacker can see the screen from an angle almost similar to that of the user, so he or she can see the character or number entered in the correct direction. Thus, at this time, there is the highest attack potential, and in the angle between 90 and 135° (225 and 270°), the attacker recognizes the character or number entered on the screen inclined 90° or more in the opposite direction, so it is more difficult for him or her to recognize the character or number as compared with at 45°. Thus, there is moderate attack potential in the relevant angle. Lastly, in the angle between 135 and 180° (180 and 225°), everything looks upside down when the attacker sees the screen, so it is difficult to judge the information entered within a short time, and thus this angle is judged to have the lowest attack potential.

Table 5. Attack potential about angle of screen

Screen Angle	Value
0–45 (315–360) degrees	1
135–180 (180–225) degrees	4
90–135 (225–270) degrees	7
45–90 (270–315) degrees	10

3.2.3 Field of view

The maximum bevel angle from which a person can see the contents displayed on the screen by looking normally is called the FOV (field of view) [14]. Human eyes have an FOV of 0º in the direction of the nose, 95º in the outer direction, 0º in the upper direction, and 75º in the downward direction; and more specifically, they can be expressed by horizontal vision and vertical sight. First, horizontal vision can be shown as in Fig. 2, and the central vision, the very middle part, is called binocular vision. The FOV of binocular vision is about 0º in both the left and right directions, and since reading skills tend to decrease if it exceeds the relevant angles, one should form the FOV within in the angle of binocular vision to read texts or symbols accurately.

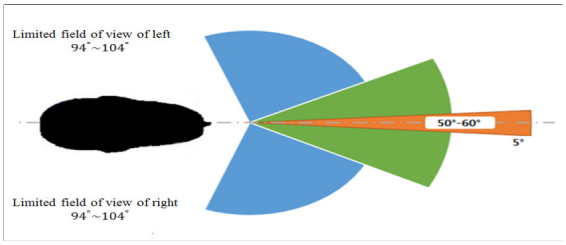


Figure 2. Horizontal FOV as seen by a man

Vertical sight can be shown as in Fig. 3, and in a seated and comfortable position for seeing, while relieving eye tension, about 15º becomes the plain sight downward, and in a standing position, 10º downward is the general sight. A total of 120 degrees, up to a maximum angle of 50º upward and up to 70º downward, are the upper and lower FOVs, respectively [15].

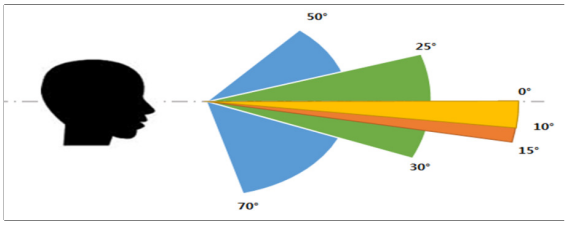


Figure 3. Vertical FOV as seen by a man

As such, a person's vision can decipher texts and symbols within the FOV range of the horizontal plane and the vertical plane. The optimal FOV is when 50 to 60º on the horizontal plane and 10 to 15º downward on the vertical plane are formed, and at this time, a sight most efficiently viewed by shoulder surfing attack is formed. Attack potential values can be shown as follows:

Table 6. Attack potential about the FOV

FOV (Field of View)		
Horizontal FOV	Vertical FOV	Value
Over 104 degrees	Over 70 degrees	1
30–104 degrees	30–70 degrees	4
0–30 degrees	15–30 degrees	7
0–5 degrees	0–15 degrees	10

3.2.4 Space and distance with legibility

The attacker should keep an appropriate distance from the target of attack and at the same time maintain the optimum distance possible for shoulder surfing attack. This should be done considering personal space and ability to recognize characters (their legibility).

3.2.4.1 Personal space

A personal space refers to the one that a person considers unconsciously to be his or her own domain. Most people give value to their own personal space, and when someone encroaches on it, they feel psychological discomfort, get angry, or alert the other person. This personal space includes intimate distance, personal distance, social distance, and public distance, shown as Fig. 4 below [16]. The intimate distance refers to the one between lovers, children, and parents; the

personal distance, to the relationship between close friends; social distance, to the practical human relationship; and public distance, to the one in speeches or lectures. In shoulder surfing attack, the most important position and distance between the user and the attacker are determined within the range of the personal space that the user recognizes.

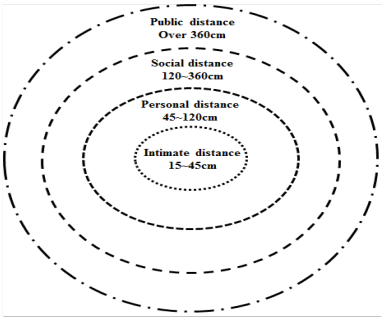


Figure 4. Personal space

Thus, when the user and the attacker are in the intimate distance (the closest distance), text legibility becomes optimum and the highest attack potential can be formed. In contrast, if they are located at the farthest distance (the public distance) the lowest attack potential may be formed, and the value of attack potential can be given accordingly.

Table 7. Attack potential about personal space

Personal space	Value
Public distance	1
Social distance	4
Personal distance	7
Intimate distance	10

3.2.4.2 Text legibility

People fix their eyes, looking at a particular thing or object for a short time, and recognize or perceive the form through that. The measure of recognition of a letter or character is called legibility, and they have the ability to read legible characters or numbers through their gaze. A monitor is usually farther than the normal reading distance (30–35 cm), so the size of characters should be larger. It should be at least 10 pt., and from 11 pt. to 14 pt. is the size that the user can read comfortably [17]. Even the same font may have different legibility and preference depending on its size, and physical conditions such as sight may affect it, so text legibility should consider various complex factors. Other factors that may affect legibility include font, text size, background color, text color, method of presentation, illuminance, sight, age, and text interval [20]. Of these, the minimum size of a character according to sight distance is shown in Table 8, and when it meets the minimum size at each distance, inconvenience of legibility reaches the minimum. With regard to text size depending on sight, as shown in a comparative study of legibility in which there was the biggest difference in the minimum text size about 15 pt when there is a difference about 4 times, the sight distance is one of the most important factors impacting legibility [21], and the sight distance may be determined within the range of the above-presented personal space. Moreover, it turned out that numbers were more legible than characters and that under bright light, the minimum legible text size was a little smaller than under dull light [22].

Table 8. Minimum text size according to sight distance

Sight Distance	Minimum Size of Character
20cm	1.4mm
30cm	2.1mm
40cm	2.8mm

In addition, the younger the subject was, the better the legibility of characters and numbers was, and when the sight distance was 50 cm and 200 cm, appropriate legible characters were 8 pt. to 22 pt. for those in their 20s and 14 pt. to 32 pt. for those in their 60s [20]. The font size was related to the screen size, and when it is assumed that the font size of the qwerty keypad provided in the current mobile banking applications is approximately 9 to 10 pt, characters in the same font size looked larger in proportion to the screen size in the following order: Galaxy Note3 (5.68 inch), Galaxy S5 (5.1 inch), Galaxy S4 (4.99 inch), and iPhone 5S (4 inch). This means that fonts on smart phones with a larger screen size looked larger than the same ones on smaller smart phones, which led to better legibility. This also means that the font size that has the greatest impact on legibility may differ depending on the smart phone device the user uses, and at the same time, using a smart phone with a larger screen size leads to better legibility, although it may be more vulnerable to shoulder surfing attack. Consequently, attack potential for legibility can be shown according to the screen size of the smart phone the user uses.

Table 9. Attack potential about legibility

Legibility		Value
Screen size	4–4.5 inches	1
	4.5–5 inches	4
	5–5.5 inches	7
	5.5–6 inches	10

3.3. Proposed Attack Potential

The proposed method adds the points of shoulder surfing attack to the existing attack potential measurement matrix since the existing measurement methods do not consider it. Thus, the classification of the rating of attack potential followed the system of the existing CEM as much as possible, as described below in this paper. The proposed attack potential can be shown as in Table 10. This added the condition of attack elements that can judge the attack potential tolerance of shoulder surfing attack to the existing attack elements. By doing so, the existing attack potential, which was not able to judge the attack potential tolerance of the shoulder surfing attack on the password input scheme, could be improved and supplemented. Moreover, it added shoulder surfing attack to the attack potential of attacks on several existing password input schemes, allowing us to demonstrate the attack potential rating of shoulder surfing attack.

This can be shown through the sum of the elements as low (0–30), medium (30–45), high (45–60), and very high (over 60), and as the sum of these values is high, the conditions that can make shoulder surfing attack succeed are met to the maximum. Using this, it can be judged that when attack potential of shoulder surfing attack falls within the “low” range at 0 to 30 points, it indicates that the probability of success in shoulder surfing attack is the lowest, while in contrast, when attack potential is over 60 points, in the range of “high or above,” it can be judged that the probability of success in shoulder surfing attack is the highest. This reflects the relative difference in vulnerability to shoulder surfing attack according to attack elements, which may be changed by the assessor, depending on the specific environment or conditions used in the attack.

Table 10. Proposed attack potential

Attack potential			
Elements			Value
Elapsed time	Within 1 day		0
	Within 1 week		1
	Within 2 weeks		2
	Within 1 month		4
	Within 2 months		7
	Within 3 months		10
	Within 4 months		13
	Within 5 months		15
	Within 6 months		17
	Over 6 months		19
Specialized knowledge	Layman		0
	Proficient		3
	Expert		6
	Complex expert		8
Knowledge about target of attack	Public information		0
	Restricted information		3
	Sensitive information		7
	Critical information		11
Period of easy exposure to attack	Unnecessary/Limitless access		0
	Easy access		1
	Moderate access		4
	Difficult access		10
Equipment	Standard equipment		0
	Specialized equipment		4
	Customized equipment		7
	Complex customized equipment		9
Perception and Recognition Time	More than 5.8 sec.		1
	4.8–5.8 sec.		4
	3.8–4.8 sec.		7
	2.8–3.8 sec.		10
Screen Angle	0–45 (315–360) degrees		1
	135–180 (180–225) degrees		4
	90–135 (225–270) degrees		7
	45–90 (270–315) degrees		10
Field of View	Over 104 degrees	Over 70 degrees	1
	30–104 degrees	30–70 degrees	4
	0–30 degrees	15–30 degrees	7
	0–5 degrees	0–15 degrees	10
Personal space	Public distance		1
	Social distance		4
	Personal distance		7
	Intimate distance		10
Legibility	4–4.5 inches		1
	4.5–5 inches		4
	5–5.5 inches		7
	5.5–6 inches		10

As a result, through the proposed attack potential, the attack potential rating to which shoulder surfing attack has been added can be judged along with the existing general password attack potential rating. If the proposed attack potential is applied to the qwerty keypad of the mobile banking applications in the market, the attack potential of shoulder surfing attack can be quantitatively known and accordingly, the attack potential rating can be known. Through relevant rating, the fact that the mobile banking applications of financial institutions are relatively more vulnerable to shoulder surfing attack can be understood. For this, the parts vulnerable to shoulder surfing attack, such as user feedback and the feedback offer time dealt with in the following

section, should be supplemented and improved upon or a picture-based password with excellent user convenience and security should be developed so that users can be safe from several types of password attack including shoulder surfing attack. In addition, the security of the applications that can perform finance-related operations, such as mobile banking applications, should be higher than for other applications, and as well, their users' security consciousness should be higher. For this, the attack techniques that could not be expressed quantitatively, such as shoulder surfing attack, can be shown using exact figures, through which users can understand which security keypads provided by mobile banking applications of financial institutions are safe from shoulder surfing attack, and thus security consciousness about shoulder surfing attack can be perceived and improved.

Table 11. Vulnerability rating of the proposed attack potential

Range of Value	Attack Potential
0–30	Low
30–45	Medium
45–60	High
Over 60	Very high

4. ANALYSIS OF VULNERABILITY BY SECURITY KEYPAD

4.1 Analysis of Secure Keypad Vulnerability

This section will analyze the safety of security keypads—the qwerty keypad and the number keypad—provided by current mobile banking applications in South Korea, and determine and analyze weak points that facilitate shoulder surfing attack. Currently, most major financial institutions (banks, securities, insurance companies) ask the user to enter a password for an account and certificate or important information such as a security care number whenever he or she carries out financial tasks using a mobile banking application. Both the qwerty keypad and number keypad attempt to achieve security through random keypad layout, but random layout with small probability values and user feedback for confirming the password the user entered have security problems vulnerable to shoulder surfing attack.

4.1.1 Random Layout

4.1.1.1 Random layout of qwerty keypad

Figure 5. Qwerty keypads provided by the mobile banking applications of financial institutions in the Republic of Korea

In entering the certificate password on mobile banking applications, the user uses a security keypad, the qwerty keypad, and they generate 1 to 2 blanks randomly on each line of the qwerty keypad for safe input. But the number of cases of the positions in which a blank may be generated can be calculated, so the randomness of the blanks can be analyzed in terms of probability.

Table 12. The values of probability for the key layout of a qwerty keypad [18]

	□		□		□		□		□		□		□		□		□		□		□	
First	1	100	1	10	2	20	3	30	4	40	5	50	6	60	7	70	8	80	9	90	0	100
			2	90	3	80	4	70	5	60	6	50	7	40	8	30	9	30	0	10		
Second	q	100	q	10	w	20	e	30	r	40	t	50	y	60	u	70	i	80	0	90	p	100
			w	90	c	80	r	70	t	60	y	50	u	40	i	30	o	20	p	10		
Third	a	100	a	20	a	2.2	s	6.6	d	13.3	f	22.2	j	13.3	k	6.6	l	2.2	l	20	l	100
			s	80	s	35.6	d	46.7	f	22.2	g	55.6	h	53.4	j	46.7	k	35.6	k	80		
					d	62.2	f	46.7	g	33.3	h	22.2	g	33.3	h	46.7	j	62.2				
Fourth	z	100	z	14.3	x	28.6	c	42.9	b	42.9	n	28.6	m	14.3	m	100						
			x	85.7	c	71.4	v	57.1	v	57.1	b	71.4	n	85.7								

As in Fig. 5, a keypad layout consisting of a total of 11 blanks on each line may be generated. Lines 1, 2, and 4 consist of 10 keys and 1 blank: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10; q, w, e, r, t, y, u, i, o, and p; and z, x, c, v, b, n, and m. Line 3 consists of 9 keys and 2 blanks: a, s, d, f, g, h, j, k, and l. The attacker can calculate the values of the key distribution in which each key may be located through probability analysis [18], and using the relevant probability values, he or she can infer the key in the position of the observed input by probability and learn the qwerty keypad layout to carry out a shoulder surfing attack.

4.1.1.2 Random layout of number keypad

The number keypad used mainly to enter an account password or security card number consists of a random configuration of 10 number keys so that it can be safe from attacks such as keystroke logging. Also, numbers are encrypted with an asterisk (*), so even if an attacker intercepts it midway through, he or she will not be able to recognize them since they will have been encrypted. Unlike the qwerty keypad analyzed above, the number keypad possesses high randomness, so it is difficult to infer 10 digits that consistently change.

Figure 6. Random number keypad

As a result, the number keypad is designed with high security in mind, but the 10 number keys are completely randomly mixed, so the user has to find the right number to press, which leads to the inconvenience of the process taking a longer time. This deteriorates the user convenience but improves safety for attack models such as keystroke logging and MITM attack. But just like when using 4 digits, which are short and vulnerable to peeping, it is useless for preventing attacks such as shoulder surfing attack, which obtains passwords by observing the user's direct input.

4.1.2 User feedback and time of feedback

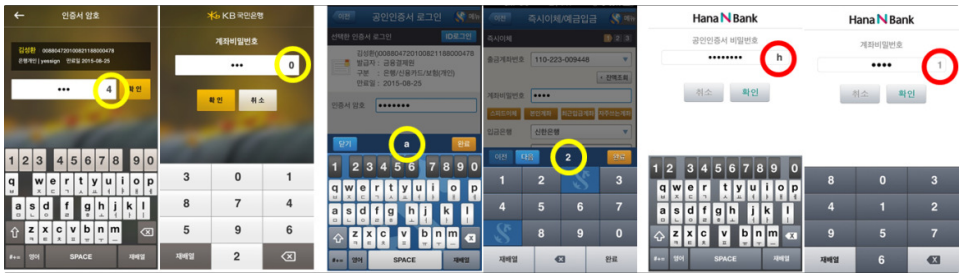


Figure 7. User feedback provided by qwerty keypads

In mobile banking applications in South Korea, security keypads have feedback processes to confirm the password the user entered, and in the processes, there are security weak points that encourage shoulder surfing attack. Both qwerty keypads and number keypads code passwords so that the keys entered are not seized by an attacker in the middle of a transmission to the server when the user touches the keypad to enter a password (that is masked using an asterisk), which is displayed before coding in order to check whether he or she entered it correctly. At this time, the suggested feedback time should consider convenience so that the user does not have any difficulty in entering and confirming his or her password, and at the same time, its security should be balanced so that it is safe from attack techniques such as shoulder surfing attack. However, since several of the security keypads of mobile banking applications continue displaying the text that the previous user entered without any time limit when they provide feedback about the values entered, the user's convenience increases, but there exists the problem that the attacker can see the text the user entered in plain text by just looking at the text provided as feedback without having to look at the key the user presses, resulting in lowering security.

Table 13. Time and method of user feedback for the mobile banking applications provided by each financial institution

	Financial Institution	Password input keypad	Encryption	Feedback Method	Feedback Time	Shoulder Surfing Attack
Mobile Banking Application	Bank A	qwerty keypad/ number keypad	O	Keep the last letter in plain language	Unlimited	Possible during plain language feedback
	Bank B	qwerty keypad/ number keypad	O	Keep the last letter in plain language	Unlimited	Possible during plain language feedback
	Bank C	qwerty keypad/ number keypad	O	Keep the last letter in plain language	Unlimited	Possible during plain language feedback
	Bank D	qwerty keypad/ number keypad	O	Keep the last letter in plain language	Limited	Possible during plain language feedback
	Bank E	qwerty keypad/ number keypad	O	Keep the last letter in plain language	Limited	Possible during plain language feedback
	Bank F	qwerty keypad/ number keypad	O	None	None	Possible while entering password

4.2 Application and Analysis of Attack Potential for Mobile Banking Security Keypad

To determine the amount of the attack potential current mobile banking security keypads have for shoulder surfing attack, based on the following attack scenario, attack potential of the above proposed shoulder surfing attack is drawn . The proposed attack scenario is composed assuming the conditions under which shoulder surfing attack can be most powerfully carried out and successful.

◇ Attack Scenario (1)
The user and the attacker located at a close distance.
The user using Galaxy Note3.
The attacker is presumed to be the user's close acquaintance.
The attacker in a low age group.
The attacker having learned random layout of qwerty keypad.
◇ Attack Scenario (2)
The user and the attacker located at a personal distance.
The user using Galaxy S5.
The attacker is presumed to be the user's close acquaintance.
The attacker in a low age group.
The attacker having learned random layout of qwerty keypad.
◇ Attack Scenario (3)
The user and the attacker located at a close distance.
The user using Galaxy S5.
The attacker is presumed to be the user's close acquaintance.
The attacker in a low age group.
The attacker having learned random layout of qwerty keypad.

The user does not have great apprehension about an acquaintance's approach to an adjacent position, so the attacker can attain the position and distance with the best legibility, and also, the user uses Galaxy Note3 with the largest screen size, so in terms of legibility, there is the optimum condition for shoulder surfing attack. In addition, through learning random layout, the attacker is familiar with the key layout, so the attacker easily recognizes and perceives them when the user uses 8 digits, the minimum requirement for a password. If values of the proposed attack potential apply, vulnerability to shoulder surfing attack can be demonstrated quantitatively. Attack potential values for mobile banking applications for each financial institution are provided in Table 14.

This is a setting of the values of attack potential according to attack scenario, and there were no differences in the values for elapsed time, specialized knowledge, knowledge about the attack target, and period to easily be exposed to attack and equipment, which are the existing elements of attack potential. And this is because the elements that might greatly affect shoulder surfing attack, such as one using a recording device, were not taken into account. However, if an attacker invests much time (elapsed time) in an attack to increase the success rate of his or her shoulder surfing attack, acquires the characteristics of the mobile banking keypad of each bank, and builds on specialized knowledge, the scores of the existing elements of attack potential may occur in several forms. So the existing elements of attack potential, also, should not be omitted.

In the proposed elements of attack potential, most attack potentials obtained the same scores, but for perception and recognition elements, there were differences in the scores, so the points were

measured differently. This is because the attacker did not have any difficulty in perceiving and recognizing the text entered by the attacker, as A, B, and C Bank's mobile banking application security keypads kept providing user feedback without any time limits. In contrast, the mobile banking security keypads of D Bank and E Bank providing user feedback for about 2 seconds and that of F Bank not providing any user feedback created more difficulty in perception and recognition than other banks that keep providing feedback.

Thus, the real attack time of the attacker may be longer than the total time the user takes to enter the password, greatly reducing the likelihood that a shoulder surfing attack will be successful. Consequently, scores of perception and recognition were measured to be lower than those of other financial institutions. Accordingly, the attack potential value of F Bank's mobile banking application was 45 points, according to attack scenario (1). It had lower attack potential than other financial institutions, so it is relatively safer. Attack Scenario (2) showed the difference by lowering personal space and legibility, the elements greatly affecting shoulder surfing attack, by one step each from Attack Scenario (1). As a result, as shown in Table 15, the values of attack potential of all banks decreased by 9 points each from the results of Attack Scenario (1), and for example, in F Bank, the rating of attack potential went down one step. Along with this, Attack Scenario (3) could lead to the result shown in Table 16, with the same personal space as Scenario (1), but with legibility lowered one step, demonstrating the importance of legibility elements in shoulder surfing attack. Thus, to compare Attack Scenarios (1) and (2), even if legibility, the most important attack element of a shoulder surfing attacker, is altered, the difference in recognition and perception by the user attack provided by the mobile banking application of each bank is an important element that determines the rating of attack potential.

As a result, for mobile banking applications to have a relatively lower attack potential from shoulder surfing attack, it is necessary to provide user feedback—which is currently provided for an infinite time—for a finite time or develop a new security keypad that provides feedback through another method to maintain usefulness and achieve security at the same time.

Table 14. Attack potential of mobile banking application for each financial company according to attack scenario(1)

Attack Elements / Mobile Banking Application	Elapsed time	Expertise	Knowledge about target of attack	Period of easy exposure to attack	Equipment	Recognition & perception	Field of View	Screen angle	Personal space	Legibility	Attack potential
Bank A	0	0	0	1	0	10	10	10	10	10	51
Bank B	0	0	0	1	0	10	10	10	10	10	51
Bank C	0	0	0	1	0	10	10	10	10	10	51
Bank D	0	0	0	1	0	7	10	10	10	10	48
Bank E	0	0	0	1	0	7	10	10	10	10	48
Bank F	0	0	0	1	0	4	10	10	10	10	45

Table 15. Attack potential of mobile banking application for each financial company according to attack scenario(2)

Attack Elements / Mobile Banking Application	Elapsed time	Expertise	Knowledge about target of attack	Period of easy exposure to attack	Equipment	Recognition & perception	Field of View	Screen angle	Personal pace	Legibility	Attack potential
Bank A	0	0	0	1	0	10	7	10	7	7	42
Bank B	0	0	0	1	0	10	7	10	7	7	42
Bank C	0	0	0	1	0	10	7	10	7	7	42
Bank D	0	0	0	1	0	7	7	10	7	7	39
Bank E	0	0	0	1	0	7	7	10	7	7	39
Bank F	0	0	0	1	0	4	7	10	7	7	36

Table 16. Attack potential of mobile banking application for each financial company according to attack scenario(3)

Attack Elements / Mobile Banking Application	Elapsed time	Expertise	Knowledge about target of attack	Period of easy exposure to attack	Equipment	Recognition & perception	Field of View	Screen angle	Personal pace	Legibility	Attack potential
Bank A	0	0	0	1	0	10	10	10	10	7	48
Bank B	0	0	0	1	0	10	10	10	10	7	48
Bank C	0	0	0	1	0	10	10	10	10	7	48
Bank D	0	0	0	1	0	7	10	10	10	7	45
Bank E	0	0	0	1	0	7	10	10	10	7	45
Bank F	0	0	0	1	0	4	10	10	10	7	42

5. CONCLUSION

In this study, we revealed attack modeling conditions appropriate for shoulder surfing attack in order to improve critical points that attack potential in the CEM cannot quantitatively express with relation to shoulder surfing attack on password input scheme. In doing so, we were able quantify and express the shoulder surfing attack that could not be quantitatively expressed by the existing attack potential.

Also, we analyzed whether qwerty keypads and number keypads, used in current mobile banking applications, would be safe from shoulder surfing attack. We carried out the analysis using existing studies and their weak points and found that providing the keyboards with a less random layout and user feedback for unlimited time would be effective in preventing shoulder surfing attack. Through this, we were able to determine concrete values and situations for the listed attack conditions of shoulder surfing attack in which shoulder surfing attack could most successfully be facilitated.

As a result, we found that the security keypad of mobile banking applications was safest from shoulder surfing attack when the vulnerability rating registered “low”; and so existing

commercial security keypads should be designed and implemented to have a minimum vulnerability rating of "low" for the proposed attack potential. Future studies should inquire into the attack potential of other password attack techniques in addition to the attack potential of security keypads of mobile banking applications by shoulder surfing attack.

ACKNOWLEDGEMENTS

This work was supported by the ICT R&D program of MSIP/IITP. [2014(10043959), Development of EAL4 level military fusion security solution for protecting against unauthorized accesses and ensuring a trusted execution environment in mobile devices]

REFERENCES

[1] Jeonghyuk Kim, Munseon Bae, and Ara Yang, "Usage of domestic Internet banking services in 2014 first quarter", Bank of Korea, May. 2014
[2] Jaesik Mun, "2014 Statistical information of the wireless communication subscriber", Ministry of Science, ICT and Future Planning, June. 2014
[3] Scott Pinzon and Kevin D. Mitnick "No Tech Hacking: A guide to Social Engineering, Dumpster Diving, and Shoulder Surfing", SYNGRESS, pp. 27-60, 2011
[4] Xiaoyuan Suo, Ying Zhu, and G. Scott. Owen, "Graphical Passwords: A Survey", IEEE, 2005
[5] L. Sobrado and J. C. irget, "Graphical Passwords", The Rutgers Scholar, An Electronic Bulletin for Undergraduate Research, vol.4, 2002
[6] S. Man, D. Hong, and M. Mathews, "A shoulder surfing resistant graphical password scheme", Proceedings of International conference on security and management, November, 2003
[7] RealUser, www.realuser.com, June, 2005
[8] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A.D. Rubin, "The Design and Analysis of Graphical Passwords", Proceedings of the 8th USENIX Security Symposium, 1999.
[9] J. Goldberg, J. Hagman, and V. Sazawal, "Doodling Our Way to Better Authentication", Proceedings of Human Factors in Computing Systems(CHI), USA, 2002.
[10] G. E. Blonder, "Graphical passwords", Lucent Technologies, Inc., Murray Hill, NJ, U. S. Patent, Ed. United States, 1996.
[11] M. N. Doja and Naveen Kumar, "User Authentication Schemes for Mobile and Handheld Services", 2007
[12] Jeongmo Lee, Eunjoo Kang, and Minsik Kim et al., "Cognitive Psychology", Hakjisa, Jan, 2009
[13] "Information Supplement: ATM Security Guidelines", PCI Security Standards Council, Jan, 2013
[14] "Field of View", Wikipedia, September, 2014
[15] F. Row, "Landscape and Visual Impact Assessments", SLD Council, 2011
[16] "Personal Space", Wikipedia, Aug, 2014
[17] Choo-Youn Chong, "Korean typography interface evaluation and development of legibility formula in smartpad device", KAIST, 2012
[18] Yunho Lee, "An Analysis on the Vulnerability of Secure Keypads for Mobile Devices", Journal of Korean Society for Internet Information, v.14, no.3, June, 2013
[19] Sooyeon Shin and Taekyoung Kwon, "STM-GOMS Model : A Security Model for Authentication Schemes in Mobile Smart Device Environments", KIISC, v.22, no.6 , Dec, 2012
[20] Inseok Lee, Seung Min Mo, Yong Ku Kong, Young Woong Song, and Myung Chul Jung, "Evaluation of Main Factors Affecting on the Legibility of One-Syllable Korean Characters and Numbers", Journal of the Ergonomics Society of Korea, Nov, 2009.
[21] Seung Min Mo, Young Woong Song, Inseok Lee, Myung Chul Jung, and Yonggu Jeong, "Legibility comparison of Korean characters and words", Ergonomics Society of Korea, May, 2009
[22] Seung Min Mo, Daemin Kim, Young Woong Song, and Myung Chul Jung, "Evaluations of Factors Affecting Legibility", Journal of the Ergonomics Society of Korea, Oct, 2008
[23] "Common Methodology for Information Technology Security Evaluation", Version 3.1, Revision 4, Sep, 2012

[24] Arash Habibi Lashkari, Samaneh Farmand, Omar Bin Zakaria, and Rosli Saleh, “Shoulder surfing attack in graphical password authentication”, International Journal of Computer Science and Information Security, Vol. 6, No.2, 2009
[25] Robert Biddle, Sonia Chiasson, and P.C. van Oorschot, “Graphical Passwords: Learning from the First Twelve Years”, ACM Computing Surveys, Feb, 2011
[26] Taekyoung Kwon, Sooyeon Shin, and Sarang Na, “Covert Attentional Shoulder Surfing: Human Aversaries Are More Powerful Than Expected”, IEEE Transactions on Systems, Man, And Cybernetics: Systems, June 2014
[27] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, and M. P. Scipioni, et al., “Back-of-device authentication on smartphones,” in Proc. CHI, 2013
[28] Joint Interpretation Library, “Application of Attack Potential to POIs”, June, 2011
[29] Joint Interpretation Library, “Application of Attack Potential to Hardware Devices with Security Boxes”, May, 2012
[30] Joint Interpretation Library, “Application of Attack Potential to Smartcards”, April, 2006
[31] Q. Yan, J. Han, Y. Li, J. Zhou, and R. H. Deng, “Designing leakageresilient password entry on touchscreen mobile devices,” in Proc. ASIACCS, 2013
[32] G. A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing information,” Psychol. Rev., vol. 63, no. 2
[33] Shiffrin, Richard; Robert Nosofsky (April 1994). "Seven plus or minus two: A commentary on capacity limitations.", Psychological Review. 2 101 (Centennial): 357–361, April 2012

AUTHORS

Sunghwan Kim received his B.S degree in Management Information System from Korea University(KU) of Korea, in 2013. He is currently working toward M.S degree in Financial Security, Korea University(KU), Korea. His research interests include Information Assurance, Financial Security, and Usable Security.

Heekyeong Noh received her B.S degree in Internet Information Engineering from Duksung Women’s Unviersity of Korea, in 2012. She is currently working toward M.S degree in Information Security, Korea University(KU), Korea. Her research interests include Password Security, Security Engineering, and CC(Common Criteria)

Chunghan Kim received his B.S degree in Computer Software Engineering from Kwangwoon University of Korea, in 2013. He is currently working toward M.S degree in Financial Security, Korea University, Korea. His research interests include Financial Security, Usable Security and Network Security.

Seungjoo Kim received his B.S., M.S. and Ph.D. from Sungkyunkwan University (SKKU) of Korea, in 1994, 1996 and 1999, respectively. Prior to joining the faculty at Korea University (KU) in 2011, He served as Assistant & Associate Professor at SKKU for 7 years. Before that, He served as Director of the Cryptographic Technology Team and the (CC-based) IT Security Evaluation Team of the Korea Internet & Security Agency (KISA) for 5 years. He is currently a Professor in the Graduate School of Information Security at KU, and a member of KU's Center for Information Security Technologies (CIST). Also, He is a Founder and Advisory Director of a hacker group, HARU and an international security & hacking conference, SECUINSIDE. Prof. Seungjoo Kim's research interests are mainly on cryptography, Cyber-Physical Security, IoT Security, and HCI Security. He is a corresponding author.

HOW TO DETECT MIDDLEBOXES: GUIDELINES ON A METHODOLOGY

Vahab Pournaghshband[1], Sepideh Hashemzadeh[2] and Peter Reiher[3]

[1]Computer Science Department, California State University, Northridge, USA
vahab@csun.edu
[2]IEEE Member
hashemzadeh.s.h@ieee.org
[3]Computer Science Department, UCLA, Los Angeles, USA
reiher@cs.ucla.edu

ABSTRACT

Internet middleboxes such as VPNs, firewalls, and proxies can significantly change handling of traffic streams. They play an increasingly important role in various types of IP networks. If end hosts can detect them, these hosts can make beneficial, and in some cases, crucial improvements in security and performance But because middleboxes have widely varying behavior and effects on the traffic they handle, no single technique has been discovered that can detect all of them.

Devising a detection mechanism to detect any particular type of middlebox interference involves many design decisions and has numerous dimensions. One approach to assist with the complexity of this process is to provide a set of systematic guidelines. This paper is the first attempt to introduce a set of general guidelines (as well as the rationale behind them) to assist researchers with devising methodologies for end-hosts to detect middleboxes by the end-hosts.

The guidelines presented here take some inspiration from the previous work of other researchers using various and often ad hoc approaches. These guidelines, however, are mainly based on our own experience with research on the detection of middleboxes. To assist researchers in using these guidelines, we also provide an example of how to bring them into play for detection of network compression.

KEYWORDS

Detection, Middlebox, Guidelines

1. INTRODUCTION

Abstractly, we often assume that the Internet follows the end-to-end principle, with smart endpoints and a dumb network. However, this general picture is very different from the capabilities of the latest technologies and the actual Internet is far more complex, with the emergence and rapidly growing prevalence of middleboxes deployed at various points in the network.

Middleboxes are defined as intermediary devices which take actions other than the normal, standard functions of an IP router on the datagram path between a source host and destination host [1]. They manipulate traffic for purposes other than simple packet forwarding. In addition to routing the traffic, middleboxes can make serious changes to network flows from altering the

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 113–126, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41209

user-data to more transparent effects such as imposing additional delay on the traffic. These influences by third party middleboxes could be for malicious, security, or performance reasons.

A wide variety of middleboxes have been proposed, implemented, and deployed during the last decade [2,27,28]. Today's enterprise networks rely on a wide spectrum of specialized applications of middleboxes. Middleboxes come in many forms such as proxies, firewalls, IDS, WAN optimizers, NATs, and application gateways, and are used for various purposes including performance and security improvement and compliance. They are an integral part of today's Internet and play an important role in providing high levels of service for many applications. Recent papers have shed light on the deployment of these middleboxes [2, 3] to show their prevalence. And a recent study [4] shows that the number of different middleboxes in an enterprise network often exceeds the number of routers. Trends such as proliferation of smartphones and wireless video are set to further expand the range of middlebox applications [5].

In some cases, middleboxes do indeed exert a real influence on traffic. In others, they merely act as if they are simple routers. Knowing the presence of middleboxes is most critical in the former case, when they are actually doing something to the traffic. This is also the easier case to detect, since a middlebox that does nothing leaves no traces of its presence. We concentrate on this more important case.

Knowing the existence of the influence of middleboxes could be beneficial to the end-hosts. Sometimes the end-hosts would behave differently based on what they sense is happening to their traffic. In such cases, an accurate detection of what is happening to that traffic is the first step. Here, to illustrate this idea, we present a number of scenarios from different categories.

Scenario I

Assume a sender is about to send sensitive data, making encryption necessary. In this case, the sender will check to determine if a strong end-to-end encryption (VPN) on the path is deployed. If he detects that strong encryption is already in place, to save energy and resources, he might choose not to encrypt the traffic stream, since encryption is a relatively expensive operation.

Scenario II

The sender detects that the receiver is using a wireless connection, but is unsure if that connection is secure. If he detects that the last link is unencrypted, he would either refrain from sending sensitive information or would apply end-to-end encryption to the channel. For example, Amazon does not provide end-to-end SSL encryption to its users who are not logged in and does not require them to log in until they are about to make the payment. This is perhaps due to lack of available resources required to encrypt all users' contents for all users. Amazon servers, to use their resources effectively while protecting users' privacy from profiling, could first sense whether the user is using a secure wireless connection or not, and then apply end-to-end encryption only if the user needs that protection. Conversely, the receiver would mark the incoming data as untrusted if he detects that the sender's wireless link is insecure.

Scenario III

An Internet user in an oppressive country might detect Internet censorship imposed by his ISP and then chooses to use a proxy to bypass it. In a different scenario, an Internet user detects wiretapping on his network and uses evasive techniques such as Tor or a VPN.
Devising a detection mechanism for middleboxes can be difficult. For instance, detecting a third party middlebox that does not alter the user data, from the end hosts' point of view, is particularly

challenging, since the effect of middleboxes of this class seems transparent to the end hosts. This is true if either the third party undoes the changes made to the data before passing it to the receiver (e.g., VPN gateways or link-layer compression) or it does not alter the data at all and affects the traffic stream similarly to normal network variation (e.g., the delay-attack [6] by a compromised node in sensor networks or the Shrew attack [7] that selectively drops packets).

There have been numerous efforts to detect various types of middleboxes in the past [8, 9, 10, 11]. However, the proposed approaches are ad hoc and mostly designed to detect only specific types of middleboxes. Despite these differences, nevertheless, there are common elements in the process leading to the design of such methods. These common elements could be identified and summarized into general guidelines.

The ultimate objective of this paper is to assist researchers who intend to conduct research focused on detection of the interference of middleboxes and introduce them to the potential challenges they might face in the process. In addition, we present some recommendations on how to overcome those challenges in certain situations. In other words, the desirable outcome we seek here is to assist with devising an accurate detection mechanism in an efficient manner with the help of systematic guidelines that have been drawn from past experiences. To the best of our knowledge this is the first attempt to devise systematic guidelines for the purpose of assisting other researchers.

While introducing the phases and steps of our proposed guidelines, we demonstrate each by applying it to an example: end-to-end detection of network compression on the path. The network compression detection approach and the corresponding results have been presented in [12] as part of our prior work on this subject.

Detecting Network Compression: A Case Study

One way to increase network throughput is to compress data that is being transmitted. Network compression may happen at different network layers and in different forms: application layer, TCP/IP header [13], IP payload [14], and link layer [15, 16].

Except for application-layer compression, compression happens at intermediate nodes, often without the knowledge of end-users. For example, in January 2013, a researcher discovered that Nokia had been applying compression to its users' data without their knowledge [16]. In this case, the intermediary was surreptitiously decrypting and re-encrypting user packets in order to effectively apply compression. Users surely would have preferred to know that this was happening, both because of the security risk and because it would render their own application-level compression unnecessary.

However, performing compression and decompression requires many resources at the intermediate nodes, and the resulting overhead can overload the intermediary's queue, causing delay and packet losses. Further, not all commercial routers come with compression capabilities [17]. Thus, some intermediaries apply compression, some do not, and generally they do not tell end-users whether they do. While managing resources effectively at end-hosts is not as crucial as it is at routers, it is still beneficial—particularly for mobile devices where resources are limited. Wasting these resources on redundant compression is undesirable. End-hosts can benefit from recognizing when compression has already being applied on a network connection.
Ideally, end-hosts and intermediaries should coordinate their compression decision, but practical problems make that ideal unlikely. Therefore, since the end-hosts have the greatest interest in proper compression choices for their data, they could detect if intermediate compression is present and adjust their behavior accordingly. An end-to-end approach to detect compression by

intermediaries can help to save end-host's resources by not compressing data when intermediaries are already doing so.

The remainder of the paper is organized as follows: Section 2 presents the guidelines for a detection methodology, followed by related work in Section 3. Section 4 concludes the paper.

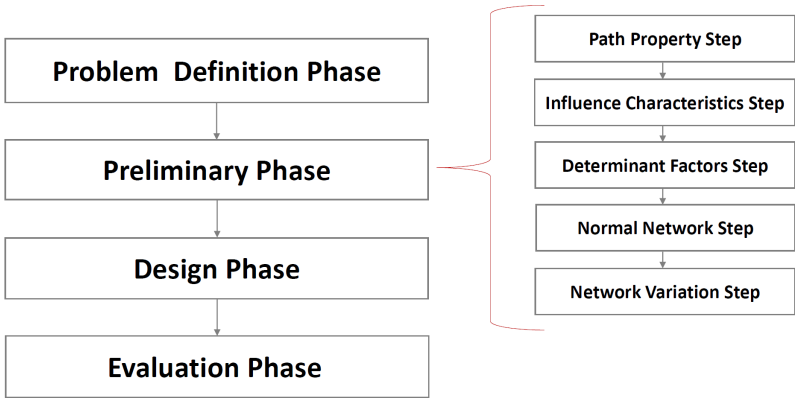


Figure 1: Phases and steps in the guideline.

2. GUIDELINES ON A DETECTION METHODOLOGY

In this section; we present the guidelines for a detection methodology. Later in this section, we show that a careful completion of the preliminary phase will lead to a significant reduction of the complexity in the design phase.

Our proposed guidelines consists of phases in a sequential order. Figure 1 illustrates the overview of the phases. The phases introduced in this guideline are:

- Phase 1: Problem Definition Phase
- Phase 2: Preliminary Phase
- Phase 3: Design Phase
- Phase 4: Evaluation Phase

In the problem definition phase we clearly define the problem and its scope by setting the assumptions. The preliminary phase consists of five steps to be followed in the presented order. The design phase in this process is when, with the help of information obtained in the preliminary phase, the detection algorithm is developed. The evaluation takes place when we validate the proposed detection mechanism devised in the design phase.

2.1 Problem Definition Phase

2.1.1 What is it that we want to detect?

What it is that we want to detect should be clearly stated. For instance, the running example we use in this paper is that we want to detect the presence of network compression on the path between two end-hosts.

2.1.2 Assumptions

The scope of the general detection problem is very broad. Many variations of the problem can be formulated based on constraints and limitations imposed on the problem: e.g., the method used for detection and the degree of detection for the problem. Before we begin with the design of any detection mechanism, we should prepare a list of assumptions we make in order to define the scope of the problem we aim to solve.

2.1.2.1 Degree of detection

A problem can be formulated to only answer an existential question about whether the network flow is influenced by some third-party or not. An instance of this problem can attempt to further characterize the influence to discover exactly what the third party is doing to the network flow.

2.1.2.2 Locating the third-party node/link

The problem statement can be phrased in several ways from determining whether the path is influenced or not to precisely locating the third-party on the path or the link(s) influenced by it.

2.1.2.3 Method of detection

There are essentially two methods of network measurements for detection available to the end-hosts: active and passive measurements. Passive measurements have the goal of minimally affecting the measured network, by merely monitoring traffic on the network and inferring measurements from the observed traffic. Active measurements involve interacting with the network to make measurements, usually by sending probe packets. In active measurements, we refer to the two end-hosts as the sender and the receiver. Based on how the receiver cooperates (if it does) in the detection process, we present three variations to this problem:

A. Receiver is uncooperative
The receiver does not respond to the sender's requests that are beyond their primary purpose of communication. For examples, most web servers expect only HTTP requests and responses.

B. Receiver is responsive
The receiver responds to the sender's requests beyond their primary purpose of communication as long as it does not require any changes on the receiver's machine. For example, the receiver responds to the sender's ICMP requests.

C. Receiver is cooperative
The receiver is willing to make necessary changes on its machine or system to fully cooperate with the sender in the detection process.

2.1.2.4 Assistance from intermediaries or other parties

If all or some intermediaries on the path are responsive, active measures can also be used to get intermediaries to respond with valuable information. In another instance of the problem, the end-hosts assume that the intermediaries are uncooperative. Another instance of the problem is where the end-hosts make use of help from volunteer nodes on the network that are not on the path.

2.1.2.5 Detection by comparing to unperturbed channel

One instance of the problem is when end-hosts have a model of the channel in the absence of a third-party's influence (perhaps captured in the past). In this case, the presence of the third party could be determined by comparing the current network behavior to the model. The other instance of this problem, however, is when the end-hosts do not have access to such information.

2.1.2.6 Static vs. dynamic third-party middleboxes

The last instance we address here is whether the middlebox is static or dynamic. In other words does its behavior remains unchanged at the time that an end-host is trying to detect it or does the third party middlebox notices the end-host's detection attempt, and hence, evades detection by changing its behavior during the detection period. In the latter case, only a stealth detection mechanism might be effective.

For our running example, end-to-end detection of network compression, we choose the following assumptions, to define the scope of the problem:

Regarding the degree of detection, we only seek to detect the presence of the intentional influence on the network flow. This problem is not about specifically locating the third-party or the influenced link on the path connecting the end-hosts. We assume end-hosts are fully cooperative and can use active measures for detection. However, due to the end-to-end nature of our problem definition, we exclude the scenario where the intermediaries are cooperative. For instance, the end-hosts should not rely on responses from pings sent to the intermediate routers. For the same reason, we exclude the instance where the end-hosts use help from volunteer nodes on the network that are not on the path. Furthermore, our proposed problem assumes that the end-hosts do not have any model of the unperturbed channel (i.e., in absence of the third party) between the end-hosts. We also assume that the end-hosts are not just normal network users and will use their resources to any required level (with some limitation on the available resources) in the detection process. And lastly, we base our detection methodology on detecting only static middleboxes.

2.2 Preliminary Phase

We believe that when starting on the design of a detection mechanism, it is necessary to carefully find answers to the following questions before offering any detection mechanism. In this phase, answering the five questions stated below will defines the steps to complete this phase.

In every step of this phase, we ensure that in addition to clearly addressing the question, and the scope of it, we also address the following questions about it: (1) Why is having an answer for this particular question important, or at least useful in the detection process? (2) What are the potential challenges in the process of answering it?

Some of the steps in this phase require a thorough understanding of the current state of technology and available tools; some others require careful analysis and examination, and some require extensive experimentation. Here, we leave the technical details out of these steps and limit ourselves to the high level presentation of the sub-problem and the expected outcome. Clearly, if a step requires experimentation that means that even more questions must be answered−such as how exactly to implement it, what would be a suitable environment to run it on, or how to prepare the experiment environment.

1) *The Path Property Step "P":* What important pieces of information about the network path properties are available to typical end-hosts?

Similar to any detection process in the real world, special tools are needed to observe and to look for signs leading to detection. Hence, it is crucial to know what information and tools are available that are observable and measurable, by the end-hosts, in order to utilize them effectively for the detection process.

For instance, some network properties such as RTT, packet delay variation, available bandwidth, and hop count (leaving out the discussion on the measurement accuracy) can be measured by the end-hosts. On the other hand, other invaluable information such as the queue size of the routers on the path is normally not available to typical end-hosts. Technically, we are only interested in standard methods and tools available for standard computers with standard equipment. For instance, we assume that in an end-host, processing time and packet arrival time can be measured and its TCP congestion control mechanism's behavior is observable. On the other hand, we exclude using irregular methods to obtain some information. For instance, we do not consider using measuring electromagnetic emissions to detect the queue size of a router.

Throughout this paper, we refer to the set of all available network properties about the path as *P*.

Challenges:

- Understanding exactly how well the properties in *P* can be measured and how the existing tools' imprecision and potential inaccuracy in measurements would affect the accuracy of the detection mechanism.

Detecting Network Compression: The receiver can measure the arrival time of packets with precision on the order of at least micro-seconds.

2) *The Influence Characteristics Step "I":* What are the characteristics of the influence *I?*

To detect the presence of *I*, we must have a way to distinguish *I* from other types of influences. Therefore, it is necessary to find unique, indicative, and distinguishing characteristics of *I*.
This step requires gathering and understanding all, if any, publicly available and known information about the internal design of the influence in question. A thorough and perhaps creative analysis is necessary to find subtle unique characteristics about *I* that can be used to detect it. These characteristics of *I* will then be exploited to help detect the middlebox. Technically, we are mainly interested in observable and measurable effects on elements of *P* and not interested in hidden or non-measurable effects of *I*. For instance, one characteristic of gateway VPNs is the relatively constant delay they impose on every packet due to the encryption/decryption time.

Challenges:

- These characteristics are not always obvious and might require a thorough and careful examination to find them.

Detecting Network Compression: If compression is placed on the bottleneck of the path, then end-hosts would potentially sense a higher bandwidth when data with lower entropy is sent. In practice, this is the primary, if not the only, objective of employing network compression.

3) *The Determinant Factors Step "D":* What elements in *P* are impacted by *I* (and to what extent)?

The values of *P*, by definition, are all that the end-hosts can know. Therefore, the only way to detect *I*, is by examining the path properties, and by looking at the changes in their values in the presence and absence of *I*.

This is an important step in identifying the factors involved in detecting *I*. We define *D* as the set of all indicative elements of *P* in detecting *I* ($D \subseteq P$). Intuitively, if the value of a path property remains constant in the presence or absence of *I*, then it is not a helpful piece of information in the detection process. It is also important to determine which of these elements are more indicative than the others in detecting *I*. Furthermore, any interdependency relationship between members of *D* must be investigated.

Theoretical analysis could lead to a hypothesis on *D*, where further experiments could verify it. Another approach to this step could be the use of network simulations. One could simulate the effects of *I* in a clean environment (i.e., in the absence of network variation and any types of traffic other than the one generated by the detection probes) and look for changes in values of various path properties in *P*.

Challenges:

- How exactly to assess the relative level of importance of $d \in D$ in detecting *I*?
- How do inaccuracies in measurements influence our findings?
- How exactly to derive the interdependency relationship between members of *D*?

Detecting Network Compression: Network compression is designed to improve, and hence should affect, the available bandwidth. Therefore, the available bandwidth is a strong candidate as a determinant factor that is potentially influenced by network compression.

4) *The Normal Network Step "N":* What are the sufficient, yet unavoidable, assumptions about the normal network behavior, in the particular network environment, required to make any comments on detectability of *I* feasible?

Even if we make the assumption that the end-hosts have no access to information on the unperturbed channel, for any detection mechanism to work, there should be a precise definition for the notion of normal behavior to use as a reference point. This is where we define the specific network environment (either in high level properties or low level). For instance, if it is multi-hop wireless network, then we expect a higher rate packet loss, whereas in a wired network, random losses are rare events. Another approach is to impose an upper-bound on the value of RTT.

These are conditions, assumptions, and constraints on the elements of *D*. But only those that are in *D* are significant, because *P* - *D*, by definition, is expected to remain relatively constant. Hence, there is no need to check whether they have deviated from normal behavior.

Basically, the violation of assumptions on the elements of *D* would be used to indicate the existence of deviations from normal network behavior.

Challenges:

- What is precisely "normal" and how do we define it?

- It is crucial to come up with not only correct, but tight assumptions about normal network behavior. Failing to arrive at correct assumptions would lead to false positives. On the other hand, loose assumptions almost certainly lead to undesirable false negatives.

Detecting Network Compression: In a normal network (i.e., in the absence of compression on the path), all packets of the same size are treated equally regardless of the data entropy of the content of their payloads.

5) *The Network Variation Step "V":* Does normal network variations (e.g., network congestion, load balancing effects, link failures, and dynamic routing effects) influence the end-to-end detection of *I*?

Any proposed detection mechanism should work in a real network. An approach that only works in a controlled and isolated environment is not very useful.

If the answer is "no" to the question posed in this step, then we can completely skip this step. For instance, detecting a third party that sends out spoofed control packets is not affected by normal network variation.

However, if the answer is "yes" (which is the case for all delay and loss-based influences), then we proceed to the following questions.

- What is the specific set of normal network variation that we care about in the detection process, *V*? For instance, one could focus only on network congestion due to its popularity.
- How well is *I* distinguishable from effects of *V*?

Note that it is generally true that congestion usually hinders the detectability of loss or delay-based influences. Clearly, if congestion, in some cases, helps the detection process, one could impose network congestion to makes detectability easier.

Challenges:

- Network variations that resemble normal loss and delay in the network are particularly challenging. This is because intentional and normal loss or delay in those cases are perhaps not easily distinguishable. This leads to another issue: how is *I* distinguishable from *V*?
- Active probing used for the detection process may contribute to network congestion.

Detecting Network Compression: Congestion influences the available bandwidth.

2.3 Design Phase

This is a crucial phase in the process where we use the information obtained in the preliminary phase to produce the detection algorithm. The design phase requires creativity and knowledge of the subject to use the information gathered in the previous phases. We recommend testing the initial approach in a simulated environment to verify the approach. Then information from Step *V* could be incorporated to produce an approach that works well in real networks. Again, we emphasize that the goal is to detect the middlebox's interference or influence on the traffic flow and if it is not interfering with the network, whether it is detectable or not is not an objective of this paper. Also, note that $\forall d \in D$ *is* used in the detection mechanism, and inaccuracies involved in measuring each element *d* must be taken into consideration.

Using the example of detecting network compression we show how the information obtained from the previous phases could lead to the design of an approach that to detects this particular influence.

Detecting Network Compression: To detect if compression is provided on the network we exploit the unique effects of compression on network flows. Assuming the original packets are of the same size, compressed low entropy data packets are expected to be considerably smaller than compressed packets containing high entropy data (the sensed bandwidth is vastly different), which in turn leads to a shorter transmission delay. Based on these facts, the sketch of our approach is as follows:

Send a train of fixed-size packets back-to-back with payloads consisting of only low entropy data. Then send a similar train of packets, except these payloads contain high entropy data instead. The receiver then measures the arrival times of the first and the last packet in the train, independently for low entropy (t_{L_1} and t_{L_N}, where N is the number of packets in a single train) and high entropy (t_{H_1} and t_{H_N}) packet trains. Note from step *P* of the preliminary phase that the typical machines are capable of measuring the arrival time of packets with a precision on the order of at least micro-seconds. Since the number of packets in the two trains is known, and all of the packets have the same uncompressed size, the following inequality will hold if some kind of a network compression is performed on the path:

$$\Delta t_L = t_{L_N} - t_{L_1} < \Delta t_H = t_{H_N} - t_{H_1}$$

This inequality suggests that the total set of highly compressible low entropy packets gets to the destination faster than the set of less compressible high entropy packets since the available bandwidth is better utilized by compression. Conversely, if the packets are not being compressed by any intermediary, then the two sides of the inequality should be almost equal. This suggests that a threshold should be specified to distinguish effects of compression from normal Internet variabilities:

$$\Delta t_H - \Delta t_L > \tau$$

The underlying rationale behind this approach is that because of the presence of network compression on the bottleneck link, the receiving party should sense a relatively higher bandwidth when the train of low entropy data is sent. This is because the same amount of data is received in both cases, but in a significantly shorter period of time when low entropy data is used.

2.4 Evaluation Phase:

Any proposed detection mechanism must be not only validated, but also evaluated under certain or all network conditions to confirm that it successfully detects the influence in question. This requires answers for questions such as:

1) How can we validate/evaluate our proposed detection algorithm?
2) How confident are we in the results of the evaluation system?
3) What metrics should we use to evaluate our detection mechanism?
4) Does the testbed used to validate our proposed detection mechanism have any shortcomings that would impact the accuracy of evaluation of our detection mechanism?

As an example, let's examine perhaps the most popular testbed used in the research community for Internet measurements: PlanetLab [18]. PlanetLab is widely used by the research community

and is generally a suitable candidate for such purposes. For every researcher working on PlanetLab there is a need to know what bias the system introduces in the collected experiment results. For instance, one should keep in mind that PlanetLab is not completely representative of the current Internet. This testbed is all about small, long running services in specific locations. There are also some shortcomings associated in the use of PlanetLab as well. For instance, we normally have very little, if any at all, control over the actual path between the chosen sender and receiver. Depending on the nature of the experiment, this could make PlanetLab somewhat ineffective.

Challenges:

- Find a platform to test the proposed detection mechanism on a global testbed, specifically one that allows us to possibly overload its nodes, if needed. It is highly desirable that the chosen testbed gives us some control over or information about the path between the nodes. Known as the ground truth problem, this is a fundamental challenge in validating detection mechanisms and the majority of previous research has acknowledged this problem to some extent. For instance, how are we going to evaluate our compression detection mechanism if we do not know whether link compression exists on the path between two nodes of the testbed?
- What metrics should be used in the evaluation process and how do we accurately measure them: e.g., false positive rate, detection rate, performance overhead, and packet delivery rate?
- Is there a need for a general evaluation system that applies to detecting all influences? How feasible is this idea?
- Recent studies, suggests that testbed results for Internet systems do not always extend to the targeted deployment. For example, Ledie et al. [19] and Agarwal et al. [20] show that network positioning systems perform much worse “in the wild” than in PlanetLab deployment. Identifying such inconsistencies to avoid false claims is not trivial.

3. RELATED WORK

While our work is the first attempt to introduce general guidelines for devising detecting methodologies for middleboxes, there has been much work in the past that proposed various approaches to detect middleboxes’ interference on traffic flows. In this section, we briefly present some of that work that has mainly security applications or implications.

Very few and even more specific works focused on detecting intentional delaying and dropping of selective packets for malicious reasons. Song et al. [5] propose two different approaches to detect and further accommodate delay attacks in wireless sensor networks. Kuzmanovic et al. [21] identify TCP victims (as the first step to detect the attacker) by monitoring their drop rates to help mitigate low-rate TCP-targeted attacks.

Other approaches introduced methodologies to detect middleboxes that send spoofed control packets on behalf of the other party. BTTest [22] focuses on detecting BitTorrent traffic blocking by ISPs that use forged TCP RST packets. Weaver et al. [23] also introduce a method to detect connection disruptions via forged TCP RST packets, but they focused only on detecting the method that China's Great Firewall uses.

FireCracker [24] proposes a framework that, through tailored probes, could be used to blindly discover a firewall policy remotely as a blackbox and without prior knowledge about the network configuration. In a more recent work, SymNet [25] proposes a static analysis technique that can model stateful middleboxes such as stateful firewalls and IDSs. There basis also been work on

detecting middleboxes that modify TCP Fields. Tracebox [26] detects middleboxes that modify TCP sequence and acknowledgement numbers, as well as TCP MSS option. Glasnost [11], on the other hand, detects modified TCP advertised window size. Detecting traffic discrimination has drawn more research attention than detecting other types of middleboxes. The majority of such detection mechanisms [8, 9, 11] use the relative discrimination technique, where they test each application (the measured flow) against a flow that is assumed to be non-discriminated (the baseline flow).

4. CONCLUDING REMARKS

In this paper we present a set of general guidelines to assist researchers with devising end-to-end detection methodologies for detecting middleboxes. More detailed, low-level guidelines could provide support for more specific design decisions. While more detailed guidelines arise left for future work, we still need to be careful not to lose the generality of the guidelines here. For instance, the evaluation phase might be generalized into a general evaluation methodology that can be used to validate all methodologies for detecting middleboxes.

We have presented a detailed case study using our proposed guidelines for detecting network compression. Evaluating the outcome of this work is difficult because it is rather a qualitative assessment. While the results are encouraging, more case studies will help to assess the effectiveness of this guideline.

REFERENCES

[1] Brian Carpenter and Scott Brim. “Middleboxes: Taxonomy and issues,” Technical report, RFC 3234, February, 2002.

[2] Justine Sherry, Shaddi Hasan, Colin Scott, Arvind Krishnamurthy, Sylvia Ratnasamy, and Vyas Sekar. “Making middleboxes someone else's problem: network processing as a cloud service,” ACM SIGCOMM Computer Communication Review, 42(4):13-24, 2012.

[3] Zhaoguang Wang, Zhiyun Qian, Qiang Xu, Zhuoqing Mao, and Ming Zhang. “An untold story of middleboxes in cellular networks,” ACM SIGCOMM Computer Communication Review, 41(4):374-385, 2011.

[4] Justine Sherry, Sylvia Ratnasamy, and Justine Sherry at “A Survey of Enterprise Middlebox Deployments,” 2012.

[5] “Enterprise Network and Data Security Spending Shows Remarkable Resilience,” http://www.reuters.com/article/2011/01/10/idUS166224+10-Jan-2011+BW20110110,January 10, 2011.

[6] Hui Song; Sencun Zhu; Guohong Cao, "Attack-resilient time synchronization for wireless sensor networks," Mobile Adhoc and Sensor Systems Conference, 2005. IEEE International Conference on, vol., no., pp.8 pp., 772, 7-7 Nov. 2005

[7] Aleksandar Kuzmanovic and Edward W. Knightly. “Low-rate TCP-targeted denial of service attacks: the shrew vs. the mice and elephants,” In Proceedings of the Conference on Applications, technologies, architectures, and protocols for computer communications, pp. 75-86, New York, NY, USA, 2003.

[8] Mukarram Bin Tariq, Murtaza Motiwala, Nick Feamster, and Mostafa Ammar. “Detecting Network Neutrality Violations with Causal Inference,” In Proc. of the CoNEXT Conference, 2009.

[9] Kanuparthy, P.; Dovrolis, C., “DiffProbe: Detecting ISP Service Discrimination,” Proceedings of IEEE INFOCOM, 2010, vol., no., pp.1,9, 14-19 March 2010

[10] Partha Kanuparthy and Constantine Dovrolis. “ShaperProbe: end-to-end detection of ISP traffic shaping using active methods,” In Proceedings of the 2011 ACM SIGCOMM conference on Internet Measurement Conference, pp. 473-482. ACM, 2011.

[11] Marcel Dischinger, Massimiliano Marcon, Saikat Guha, P Krishna Gummadi,Ratul Mahajan, and Stefan Saroiu. “Glasnost: Enabling End Users to Detect Traffic Differentiation,” In NSDI, pp. 405-418, 2010.

[12] Pournaghshband, V.; Afanasyev, A; Reiher, P., "End-to-end detection of compression of traffic flows by intermediaries," In Proceedings of IEEE Network Operations and Management Symposium (NOMS), 2014 IEEE , vol., no., pp.1,8, 5-9 May 2014.
[13] L. Jonsson, G. Pelletier, and K. Sandlund, "RFC 4995: The Robust Header Compression (ROHC) Framework," Network Working Group, pp. 1–40, 2007.
[14] A. Shacham, B. Monsour, R. Pereira, and M. Thomas, "IP Payload Compression Protocol (IPComp)," RFC 3173, 2001.
[15] R. Friend and W. Simpson, "RFC1974: PPP Stac LZS Compression Protocol," 1996.
[16] "Nokia hijacks mobile browser traffic, decrypts HTTPS data," http:www.zdnet.comnokia hijacks-mobile-browser-traffic-decryptshttps- data-7000009655, 2013.
[17] HP Support FAQs, Using Compression with HP Router Products. http://www.hp.com/rnd/support/manuals/pdf/comp.pdf.
[18] Brent Chun, David Culler, Timothy Roscoe, Andy Bavier, Larry Peterson, Mike Wawrzoniak, and Mic Bowman. "PlanetLab: an overlay testbed for broad-coverage services," SIGCOMM Computer. Communication. Rev, 33(3):3-12, July 2003.
[19] J. Ledlie, P. Gardner, and M. Seltzer. "Network coordinates in the wild," In Proc. of NSDI, volume 7, pp. 299-311, 2007.
[20] S. Agarwal and J.R. Lorch. "Matchmaking for online games and other latency-sensitive P2P systems," In ACM SIGCOMM Computer Communication Review, volume 39, pp. 315-326. ACM, 2009.
[21] Chia-Wei Chang; Seungjoon Lee; Bill Lin; Wang, J., "The Taming of the Shrew: Mitigating Low-Rate TCP-Targeted Attack," Distributed Computing Systems, 2009. ICDCS '09. 29th IEEE International Conference on, vol., no., pp.137,144, 22-26 June 2009.
[22] Marcel Dischinger, Alan Mislove, Andreas Haeberlen, and Krishna P. Gummadi. 2008. "Detecting BitTorrent blocking," In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (IMC '08). ACM, New York, NY, USA, 3-8.
[23] NicholasWeaver, Robin Sommer, and Vern Paxson. "Detecting Forged TCP Reset Packets," In NDSS, 2009.
[24] Taghrid Samak, Adel El-Atawy, and Ehab Al-Shaer. Firecracker: "A framework for inferring firewall policies using smart probing," In Network Protocols. ICNP 2007. IEEE International Conference on, pp. 294-303. 2007.
[25] Radu Stoenescu, Matei Popovici, Lorina Negreanu, and Costin Raiciu. SymNet: "static checking for stateful networks," In Proceedings of the 2013 workshop on Hot topics in middleboxes and network function virtualization, pp. 31-36., 2013.
[26] Gregory Detal, Benjamin Hesmans, Olivier Bonaventure, Yves Vanaubel, and Benoit Donnet. "Revealing Middlebox Interference with Tracebox," In Proceedings of the 2013 Conference on Internet Measurement Conference, IMC '13, pp. 1-8, 2013.
[27] Vahab Pournaghshband, Leonard Kleinrock, Peter Reiher, and Alexander. Afanasyev, "Controlling applications by managing network characteristics," In IEEE International Conference on Communications (ICC), 2012.
[28] Vahab Pournaghshband, Majid Sarrafzadeh, and Peter Reiher. "Securing legacy mobile medical devices," In Proc. of Int. Conf. Wireless Mobile Communication and Healthcare, 2012.

AUTHORS

Vahab Pournaghshband is an Assistant Professor in the Computer Science department at California State University, Northridge. He received his Ph.D. from the Computer Science department at University of California, Los Angeles (UCLA) in 2014. He received his M.Sc. in Computer Science from University of California, Berkeley. Also from UC Berkeley, he received his B.Sc. in Electrical Engineering and Computer Science. Dr. Pournaghshband has done research in the fields of computer networks, computer security, and computing education.

Sepideh Hashemzadeh received her B.S. in Engineering Science from University of Tehran in 2013. She has done research in the fields of embedded systems and computer networks. Her research interests focus on security aspect of embedded systems and computer networks.

Peter Reiher received his B.S. in Electrical Engineering and Computer Science from the University of Notre Dame in 1979. He received his M.S. and Ph.D. in Computer Science from UCLA in 1984 and 1987, respectively. He has done research in the fields of distributed operating systems, network and distributed systems security, file systems, ubiquitous computing, mobile computing, and optimistic parallel discrete event simulation. Dr. Reiher is an Adjunct Professor in the Computer Science Department at UCLA.

SECURE TRANSMISSION IN WIRELESS SENSOR NETWORKS DATA USING LINEAR KOLMOGOROV WATERMARKING TECHNIQUE

Bambang Harjito[1] and Vidyasagar Potdar[2]

[1]Department of informatics, Mathematics and Natural Science Faculty, Sebelas Maret University, Surakarta, Indonesia
harjitob2011@gmail.com
[2]School of Information System, Curtin University, Perth, Australia
v.potdar@curtin.edu.au

ABSTRACT

In Wireless sensor networks (WSNs), All communications between different nodes are sent out in a broadcast fashion. These networks are used in a variety of applications including military, environmental, and smart spaces. Sensors are susceptible to various types of attack, such as data modification, data insertion and deletion, or even physical capture and sensor replacement. Hence security becomes important issue in WSNs. However given the fact that sensors are resources constrained, hence the traditional intensive security algorithms are not well suited for WSNs. This makes traditional security techniques, based on data encryption, not very suitable for WSNs. This paper proposes Linear Kolmogorov watermarking technique for secure data communication in WSNs. We provide a security analysis to show the robustness of the proposed techniques against various types of attacks. This technique is robust against data deletion, packet replication and Sybil attacks

KEYWORDS

Linear Feedback shift Register, Digital watermarking technique and Wireless Sensor Networks

1. INTRODUCTION

Wireless Sensor Networks (WSNs) have the capability for sensing, processing and wireless communication all built into a tiny embedded device [1]. This type of network has drawn increasing interest in the research community over the last few years. This is driven by theoretical and practical problems in embedded operating systems, network protocols, wireless communications and distributed signal processing. The primary function of WSNs is to collect and disseminate critical data that characterize the physical phenomena within the target area.

We know that WSN nodes have low power supply and limited computational capability because they operate on batteries. Given their limited power supply it becomes challenges to use store for ensuring security. There are numerous security dimensions like authenticity, integrity, copyright data protection. Watermarking techniques are been investigated for addressing some of these issues like tampering, data authentication, copyright and detection etc. Watermarking algorithms are shown to be less energy demanding and the recent literature shows that incorporate

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 127–146, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41210

watermarking in WSN is feasible. Hence the research in the area of watermarking and WSN is becoming increasingly important. Watermarking technique is a lightweight technique that was used traditionally for providing copyright protection for multimedia data like images and video clips. Watermarking algorithms are much lighter and require less battery power and processing capabilities than cryptographic-based algorithms. Another advantage for the watermarking-based algorithms is that the watermark is embedded directly into the sensor data; there is no increase in the payload. While cryptography provides no protection after the content is decrypted, watermarking provides protection in secrecy at all times because the watermark is an inseparable constituent part of the host media [6-8]. Hence the research in the area of watermarking and WSN is becoming increasingly important. With the concept of cyber physical system, i.e., on web of things this research is becoming main stream and the importance of this research has become even more significant. The objective of this paper we present on secure data trnasmittion in WSNs using watermarking technique.

2. RELATED WORKS

In the last few years, there are many researches who studies on digital watermarking technique for normal data types for example texts, images, audios, videos. and even relational databases [2-4] But there are only a few research works on digital watermarking techniques for WNSs [5] [6, 7]. Feng, J.P et.al [5] developed the first system of watermarking technique to embed crypto logically encoded authorship signatures into data and information acquired by wireless embedded sensor networks. Sion et.al [6] provided copyright protection to data stream owners and authorized users. Consider the case where a stream is generated and safely transmitted from the sensors to the base station. A watermark is applied to the stream at the base station. The data are then transmitted to an authorized user. The owner and authorized users need a way to show that the data were generated by them and they want to prove that the stream was illegally obtained by the attacker. One commonly accepted way to prove ownership is the use of embedded watermarks. This technique works by embedding a watermark bit into major extremes, which are extremes that will survive any uniform sampling. F. Koushanfar et.al [8] present an active watermarking technique that can be used on the data that is processing during the common sensor fusion application from sensor of different modalities. Xiaoet.al [9]proposed a watermarking technique for protecting copyright by taking advantage of the characteristic of the sending time. Based on digital watermarking, Zhang, W, et.al [10] proposed an end-to-end, watermark statistical approach for data authentication that provides inherent support for in-network processing. In this technique authentication information is modulated as watermark and then is embedded to the sensory data at the sensor nodes. Communication protocol for WSN is introduced by Xuejun R et.al [11] for sensitive data transmission. The technique use sensitive information as watermark. The watermark is then embedded into sensory data in the sensor nodes. A threshold is used for avoiding the alteration of the lowest to make a big influence to sensory data's precision. Kamel et.al [12] introduced a technique for providing data integrity. This technique based on distortion free watermarking embeds the watermark in the order the data element so that it does not cause distortion to the data.

Usually there are two main purposes for watermarking. One of is to protect the copyright of the author. The other is to provide data integrity and to do authentication by using user's identity as watermark information. Compared with authentication schemes based on public key ciphers, the watermarking based authentication has the advantages of lower computational complexity and being invisible to adversaries. In fact, besides these purposes, watermarking technique can also be used to transmit some secret information through unsecure channels without encryption. The table1 shows their approach and their purpose many researchers who work on the watermarking technique for WSNs.

Although some research works attempted to apply digital watermarking technique into wireless sensor networks for copyright protection, authentication and integrity purposes, most of existing studies were only limited to secure data communication. No watermarking based secure data communication method has been found in related works. Therefore the purpose of this paper is that it presents secure data transmission in WSNs using watermarking technique

Table 1 Watermark embedding approaches and their purpose

Author	Watermark embedding technique	Purpose
Feng et al. [5]	Adding watermark constraint to processing step during network operation	Copyright protection
Sion et al. [6]	Selection criteria using MSB	Copyright protection
Koushanfar et al. [8]	Adding watermark constraint to processing step during network operation	Copyright protection
Xiao et al. [9]	By modification the embedding bit of each packet. LSB	Copyright protection
Zhang et al. [10]	The watermark sensory data, $d(x,y) = w(x,y)+o(x,y)$, w(x,y) is the watermark for sensor node and O(x,y) is sensory data	Authentication
Xuejun et al. [11]	IIS = input integer stream, IBs=input binary stream. T = Threshold, If IIS $\geq$T "IBS=1" become "IBS=0" Else "IBS=0" become "IBS=0"	Authentication
Kamel et al. [13]	Concatenation of the current group hash value group *gi* and next group hash value group *gi+1*. *Wi = HASH (K || gi || SN)* *SN = serial number*	Integrity

3. AN OVERVIEW OF DIGITAL WATERMARK

Watermarking technique is the process of embedding information which allows an individual to add hidden copyright notices or other verification messages to digital audio, video, or image signals and documents object [14-16]. Such hidden message is a group of bits describing information pertaining to the signal or to the author of the signal. The signal may be audio, pictures or video, for example, if the signal is copied, then the information is also carried in the copy. Watermarking seeks to embed a unique piece of data into the cover medium. The specific requirements of each watermarking technique may vary with the applications and there is no universal watermarking technique that satisfies all the requirements completely for all application.

Watermarking system as a communication task consists of three main stages: watermark generation process, watermark embedding process that including information transmission and possible attacks through the communication channel and detecting process that watermark retrieval

3.1 Watermark generation process

Generation process is the first step and a very critical of the process. The requirements of watermark generation process are unique and complexity. The watermark message contains information that must be unique such as simple text [5] [8] The key embedding is also unique in order to make a secrecy key such as binary stream [13] [17] [18] [9] [19] and pseudorandom sequence [10]. Both the watermark message and the key embedding are as input and they then are processed in the watermark generator to produce a watermark signal. Examples of the

watermark generator are Hash function [13] [5] [8] [17] [18] [19] and product function . The watermark signal is a kind of signal or pattern that can be embedded into cover medium. There two types of watermark signal, i.e., meaningful and meaningless watermark. Examples of the watermark meaningful are logo

3.2 Watermark embedding Process

Embedding process is the second step of the watermarking system. This process is undertaken by an embedder and can be done in the transform domain such as Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT). The embedder combines the cover medium, the watermark signal, the sensed data and key embedding and it then creates watermarked cover medium. Examples of the cover medium are packed data, text, image, audio signal and video. The watermarked cover medium is perceptually identical to the cover medium. The watermarked cover medium is then transmitted by the sender through the unsecure communication channel such as wireless and radio channel. During transmission, there is anything that interfere in the communication process such noise, decreasing the quality of transmitting and a watermarked cover medium dropped. The other thing is that watermark attacks such as cropping, compression, and filtering, the aim of this attack is removed the watermark signal from the watermarked cover medium

3.3 Detecting and Extracting Process

The end of the watermarking system detects or extracts process that is a crucial part because the sender can identify and provide information to the intended receiver. The detecting or extracting is undertaken by a detector. The detecting process consists of an extraction unit to first extract the watermark signal and later compare it with the cover medium or not inserted. The extracting process can be divided into two phases, locating the watermark and recovering the watermark information. There are two types detection: Informed detection and blind detection according whether the cover medium is needed or not in the detection process. For informed detection which means the cover medium such as a packet data, original image and original signal, the watermarking system is called private watermarking. For the blind detection that does not need the cover medium is used for detection, the watermarking system is called a public watermarking.

4. PROBLEM DESCRIPTION

In application the wireless sensor networks, all communication between different nodes are send in broadcast fashion through communication channel where any node become attack target with external and internal security risk including eavesdropping, leak, temper, disrupt and other. In the special application fields, if the data transmission is not reliable, the security of the whole networks will be affected. Secure data transmission between sensors nodes have become important issue because an attacker can easily eavesdrop on, inject and manipulate a sensor node. Secure data sensor networks use many cryptographic algorithms. These techniques need thousands or even millions of multiplication instructions in order to perform operations [20-24]. The essence of the public key encryption for WSNs is keeping information the plain packet data secret namely securing communication in the presence of attackers, verifying authenticity of trusted parties and maintaining transaction integrity. In the previous section, we conducted an in-depth literature survey of watermarking approach in WSNs solution and their purposes and we identified that only limited presented. a solution to the address of issues. This gives us the rationale to present our solution secure data transmission model based on watermarking technique

5. PROPOSED MODEL WATERMARKING TECHNIQUE

In this section, we give a general overview of our solution watermarking technique to protect the reliability of data transmissions. The secure data communication model based on watermarking is illustrated in Figure 1. According to the model, this model consists of four steps : (a) cover medium process (b) Watermark generator process (c) Embedder process (d) Detecting or extracting process. The cover medium process is the process to generate a cover medium by using an atomic trilateration process. The watermark generator process is to create watermark constraint and message sensed data. This process requires a sensed data whereas the data through the LFSR process, partitioned process and Kolmogorof rule process. The embedder process is the process to generate a cover medium watermarked and the process of detecting is to detect the watermark signal

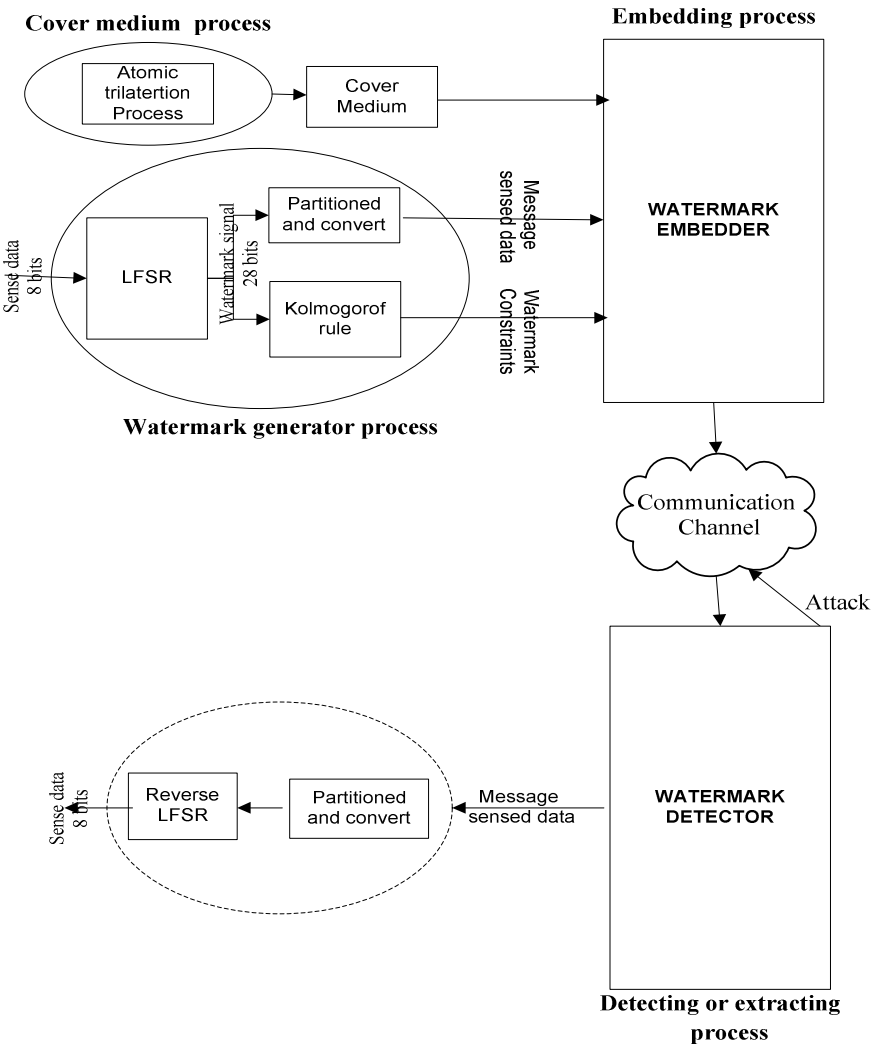


Figure 1 secure data transmission model based on watermarking

We next explain the four steps, we begin cover medium process

5.1 Generate Cover medium

In this section, we explain the process of generating cover medium by using atomic trilateration process (Pseudocode 1) With respect to a two-dimensional sensor networks, atomic trilateration is a well-known procedure by which a sensor node in a networks can determine its position by using the position of and distances to at least three other sensor nodes of know location. From these distance and position, a sensor node which is trying to determine its location can generate a nonlinear system programming.

Pseudocode 1. Generate Cover Medium

Input : $(x_A, y_A), (x_B, y_B)$, (x_C, y_c), T_c , t_{DA}, t_{DB}, t_{DC}, (x_D, y_D) , ε_t , ε_{DA} , ε_{DB} , ε_{DC} δ_1 , δ_2 , δ_3

Output: The cover medium is

$$\min f = \varepsilon_t + \varepsilon_{DA} + \varepsilon_{DB} + \varepsilon_{DC} + \delta_1 + \delta_2 + \delta_3$$

Constraints

$$\left(\sqrt{(x_D - x_A)^2 + (y_D - y_A)^2 + (z_D - z_A)}\right) - (331.4 + 0.6(T_c + \varepsilon_t))(t_{DA} + \varepsilon_{DA})) \le \delta_1$$

$$\left(\sqrt{(x_D - x_B)^2 + (y_D - y_B)^2 - (z_D - z_B)}\right) - (331.4 + 0.6(T_c + \varepsilon_t))(t_{DB} + \varepsilon_{DB})) \le \delta_2 \qquad \text{Eq (1)}$$

$$\left(\sqrt{(x_D - x_C)^2 + (y_D - y_C)^2 - (z_D - z_C)}\right) - (331.4 + 0.6(T_c + \varepsilon_t))(t_{DC} + \varepsilon_{DC})) \le \delta_3$$

Steps :

1. Compute $V_s = 331.4 + 0.6T_c$
2. Compute $d_{DA} = V_s * t_{DA}$, $d_{DB} = V_s * t_{DB}$, $d_{DC} = V_s * t_{DC}$. Where d_{DA} d_{DB} and d_{DC} is between node D and the sensor nodes are then measured using TDoA.
3. Append ε_t error of measurement time to step (2)
4. Append ε_{DA}, ε_{DB}, and ε_{DC} errors of measurement distance to step (2).
5. Compute $d_{DA} = \sqrt{(x_D - x_A)^2 + (y_D - y_A)^2}$ $d_{DB} = \sqrt{(x_D - x_B)^2 + (y_D - y_B)^2}$ $d_{DA} = \sqrt{(x_D - x_C)^2 + (y_D - y_C)^2}$
6. Append δ_1 , δ_2 and δ_3 errors between the Euclidean distances step (3)
7. Replacing d_{DA}, d_{DB} and d_{DC} from step (2) to step (3)and then compute them.
8. Print cover medium

5.2 Watermark generation process

Generation process is the first step and a very critical of the process. The requirements of watermark generation process are unique and complexity. The watermark message contains information that must be unique such as text and sensed data. The watermark key is also unique in order to make a secrecy key such as binary stream, integer and amplitude. Both the watermark message and the watermark key generator are as input and they then are processed in the watermark generator to produce a watermark signal. The process of generate watermark signal consists of five steps: (1) converting sensitive data into binary sequence, (2) Linear Feedback Shift Register (LFSR) to create watermark signal, (3) Kolmogorof rule to produce watermark constraints, (4) Partitioning and convert to decimal number from watermark signal to produce message sensed data and. Let we explain each of steps

5.2.1 Converting Sensitive data into binary sequence

The first step is that converting sensitive data into binary sequence. Any data of which the compromise with respect to confidentiality, integrity, and/or availability could have a material adverse effect on coventry interest, the conduct of agency programs. This data is called a sensitive data. The sensitive data is directly proportional to the materiality of a compromise of the data with respect to these criteria. Shih, F et.al [25] present finding sensitive data and privacy issue of applications in Body Sensor Networks(BSN). In BSN, the applications collect sensitive physiological data of the user and send to other parties for further analyses. The sensitive data are heart rate and Blood Pressure. These data are required to be protected and then these data will be converted scalar data into binary stream. WSN has gathered a blood pressure patient. The patient blood pressure is 120 so the digit sequence of 120 padded with zeros so that it is of total length 8. d= dec2bin([120],8) = 01111000

5.2.2 Generating watermark signal using LFSR

One method of forming a binary sequence for generating watermark is to apply a LFSR whose characteristic polynomial is primitive [26, 27]. LFSR is a shift register whose input bit is a linear function of its previous state. The only linear function of single bits is exclusive-or (*xor)*, therefore it is a shift register whose input bit is driven by *xor* of some bits of the overall shift register value.

LFSR can be defined by a recurrence relation:

$$s_{K+n} = \sum_{i=0}^{n-1} c_i s_{k+1}, \text{ where } k \geq 0, n \in Z \text{ and the } c_i \text{ are binary constants such that } c_O = 1. \quad \text{, Eq (2)}$$

associated with such a recurrence relation is a binary polynomial

$$f(x) = c_0 + c_1 x + \ldots + c_{k-1} x^{k-1} + x^k, \qquad \text{Eq (3)}$$

called the characteristic polynomial of the LFSR. The coefficient c_i are feedback constants. Such sequence can be mechanized by using a LFSR whose tap setting are defined by the feedback constants.

We implemented (pseudocode 2) to generate a watermark signal, we use the sensory data as the initial state of LFSR , i.e., "01111000" and the binary polynomial $f(x) = 1 + x + x^5 + x^6 + x^7$. This binary polynomial is written by [1 2 5 6] as key embedding . We then get the 28 binary sequence is 00011110 000011011100 1100 0111 .This binary sequence is called a watermark signal.

Pseudocode 2. Generate watermark signal

Input : Sensed data, coefficients c_i of the binary polynomial as watermark key

Output : 28 bits watermark signal

Steps :

1. Convert sensed data into binary sequence.
2. Use the coefficients c_i of the binary polynomial $f(x)$ as watermark key

3. Generate an infinite binary sequence using the coefficient c_i into a LFSR (s_{K+n}).
4. The infinite binary sequence cut from 1 to 28 as watermark signal.
5. Print 28 bits watermark signal

5.2.3 Kolmogorov rule to create watermark constraints

Andrew nikolaevich Kolmogorov [28] states that complexity of an object is the length of shortest computer program that can reproduce the object. The Kolmogorov complexity is defined a probability distribution under which worst-case and average-case running time are the same. We know that kolmogorof rule is the short description length of overall description interpreted by computer. The three papers [5, 29, 30] used the kolmogorov rule for numbering the variables of linear combination in the optimization objective function and a set of constrains. We also use the kolmogorov rule. This rule can be seen in Table 2

Table 2 the kolmogorov rule

1	2	3	4	5	6	7
ε_t	ε_{DA}	ε_{DB}	ε_{DC}	δ_1	δ_2	δ_3

We number (pseudocode 3) these variables by using kolmogorov rule.

Pseudocode3. Generate watermark constraints
Input : 28 bits watermark signal
Output: Watermark constraints
Steps
1. Group 28 bits watermark signal into group of 7 bits each.
2. Match the bit number with corresponding variable number from table 2.
3. If a bit one is assigned a variable with in a group that variable is included in the linear
4. Else a bit zero is assigned a variable with in a group that variable is not included in the linear.
5. Go to 2
6. Print watermark constraints

5.2.4 Partitioned and convert to create message sensed data

In this section, we explain how a message sensed data created. To create this message sensed data (pseudocode 4), 28 bits watermark binary that resulting from generating watermark signal is used.

Pseudocode 4. Generate create message data
Input : 28 bits watermark signal
Output: message sensed data
Steps:
1. Group 28 bits watermark signal into group of 4 bits each.
2. Convert each of group into decimal number to get weight factors.
3. Print message sensed data

5.3 Watermark Embedding process

Embedding process is the second step of the watermarking system that is undertaken by an watermark embedder. The embedder combines the cover medium, the watermark constraints and the message sensed data and it then creates watermarked cover medium. The watermarked cover medium is perceptually identical to the cover medium. The figure 2 shows the watermark embedding process. The watermark signal is converted to become watermark constraints by using kolmogorov rule. The watermark constraints consist of four constraints that will be added into the Equation 1. The message sensed data is also inserted into the coefficient objective of Equation 1. This message sensed data is a weight factors that obtained by partitioning the watermark signal into 7 sections. The watermark signal is converted to become watermark constraints by using kolmogorov rule. The watermark constraints consist of four constraints that will be added into the Equation 1. This message sensed data is a weight factors that obtained by partitioning the watermark signal into 7 sections. The redundant constraints are added into the Equation 1. The watermarking embedding process (Pseudocode 5) can be shown in Figure 2.

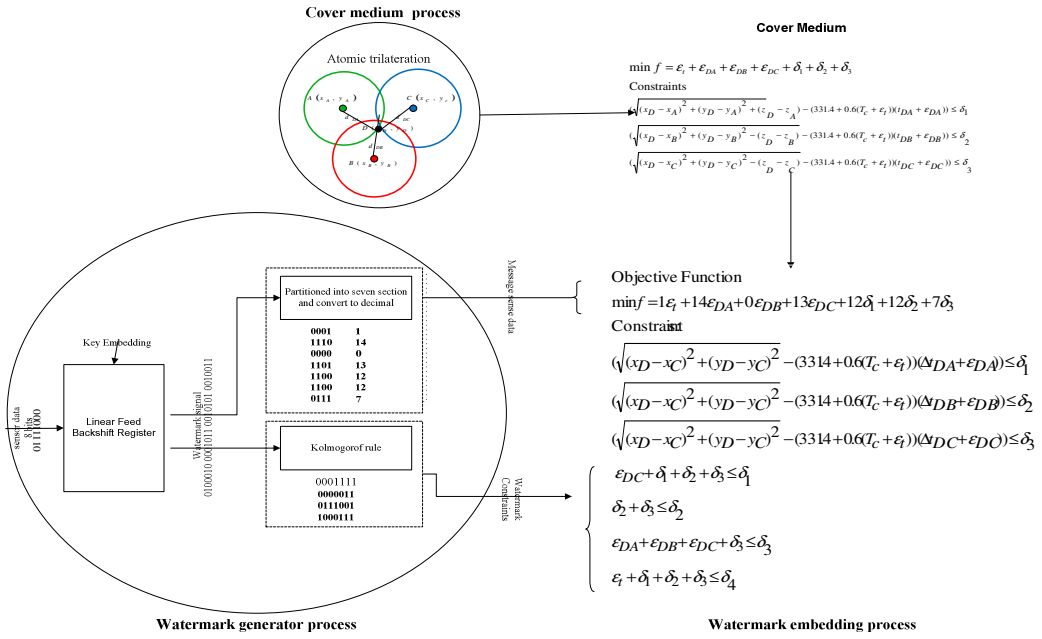


Figure 2 Watermark Embedding Process

Pseudocode 5. The process of embedding

Input: cover medium, Watermark constraints, message sensed data.

Output: (x_D, y_D), $\varepsilon_t, \varepsilon_{DA}, \varepsilon_{DB}, \varepsilon_{DC}$ $\delta_1, \delta_2, \delta_3$ and min f

Steps :

1. Generate $(x_A, y_A), (x_B, y_B)$ and (x_C, y_c), using uniform distribution on interval [0,1].
2. Generate t_{DA}, t_{DB}, and t_{DC} using uniform distribution on interval [0.02,0.1].
3. Generate δ_1, δ_2 and δ_3 using gauss distribution on interval [0,1].

4. Generate $\tau_1, \tau 2, \tau_3$ and τ_4 using gauss distribution on interval [0,1], So that these value do not harm to the feasibility of the solution of the cover medium
5. Generate τ_c using gauss distribution on interval [0,1].
6. Change coefficient objective f to weight factor of message sensed data respectively.
7. Append watermark constraints into cover medium
8. Compute and print (x_D, y_D), $\varepsilon_t, \varepsilon_{DA}$, $\varepsilon_{DB}, \varepsilon_{DC}$ δ_1, δ_2, δ_3 and min f

5.4 Watermark detecting and extracting process

The process of detecting watermark into has not yet explained in Feng Jasica P et.al [5] and F. Koushanfar et.al [8]. Both of them are only explain the process of embedding watermark. To verify the presence of the watermark, we adopt the concept of Cox *et al* [31]. Cox draw parallels between their technology and spread-spectrum communication since the watermark is spread over a set of visually important frequency components Let *X* be the error from the optimal solution without watermark and *X'* be the error form the optimal solution with watermark. For detecting the watermark, a correlation value or similarity measure is used in most of these methods. Here to verify the presence of the watermark constraints, the similarity measure between the normalized difference error from the optimal solution between the watermarked solution and the solution obtained without watermarked $C' = X' - X$. Adding the message sensed data into the Equation 1 is called the equation without watermark constraints. Adding the message sensed data and the watermark constraints are called the Equation 1with watermark. The similarity measure is given by the normalized correlation coefficient $sim(C', X') = \frac{C'.X'}{\sqrt{X'.X'}}$. Subsequently, since the expected result is dyadic (i.e. the cover medium '*is* ' or 'is not ' watermark), some kind of threshold is needed. The watermarking detecting process can be shown in Figure 3. This process (Pseudocode 6) is also can be used to obtain the value of threshold. This threshold is extracted by statistical rules and usually has a strong mathematical formulation. There are two kinds of errors in such schemes. *False-positive* corresponds to the case of detection of non-existing watermarks signal. *False-mark*

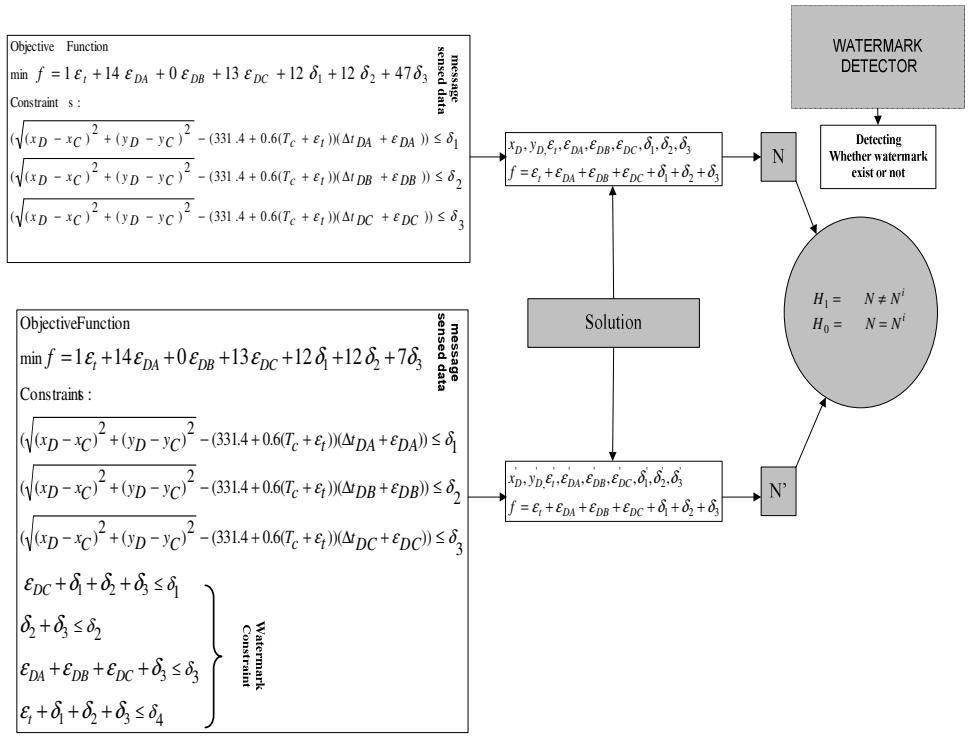


Figure 3 Watermark detecting process

stands for the case that the watermark signal exists but cannot be detected. Although well-reasoned, the existing thresholds many times lead to false-negative errors. We use False-negative to determine the watermark signal presents or not.

$H_1 = \quad N \neq N^{'}$ the cover medium is watermarked

$H_0 = \quad N = N'$ the cover medium is not watermarked

Pseudocode 6. The process of detecting

Input : x=[$\varepsilon_t, \varepsilon_{DA}, \varepsilon_{DB}, \varepsilon_{DC}\ \delta_1, \delta_2, \delta_3$], x'=[$\varepsilon_t^{'}, \varepsilon_{DA}^{'}, \varepsilon_{DB}^{'}, \varepsilon_{DC}^{'}, \delta_1^{'}, \delta_2^{'}, \delta_3^{'}$] and x"=[$\varepsilon_t^{''}, \varepsilon_{DA}^{''}, \varepsilon_{DB}^{''}, \varepsilon_{DC}^{''}, \delta_1^{''}, \delta_2^{''}, \delta_3^{''}$]

Output : Watermark signal robust or not robust

Steps :

1. Compute N = | [$\varepsilon_t, \varepsilon_{DA}, \varepsilon_{DB}, \varepsilon_{DC}\ \delta_1, \delta_2, \delta_3$]| , N' = | [$\varepsilon_t^{'}, \varepsilon_{DA}^{'}, \varepsilon_{DB}^{'}, \varepsilon_{DC}^{'}, \delta_1^{'}, \delta_2^{'}, \delta_3^{'}$]| and N'' = | [$\varepsilon_t^{''}, \varepsilon_{DA}^{''}, \varepsilon_{DB}^{''}, \varepsilon_{DC}^{''}, \delta_1^{''}, \delta_2^{''}, \delta_3^{''}$]|
2. Compute c=x'-x and c'=x"-x
3. Compute normalized correlation the results of error the cover medium without watermark constraints $treshold = \frac{C.X^{'}}{\sqrt{X^{'}.X^{'}}}$)
4. Compute normalized correlation the results of error the cover medium with watermark constraints $sim(C^{'}, X^{'}) = \frac{C^{'}.X^{'}}{\sqrt{X^{'}.X^{'}}}$.
5. If $N \neq N^i$ watermark signal exits

6. If $N = N^i$ watermark signal does not exist
7. If threshold $treshold < sim(C^{'}, X^{'})$ watermark signal is robust go to 9
8. If threshold $treshold < sim(C^{'}, X^{'})$ watermark signal is not robust
9. Algorithm the process of extracting message sensed data

The extracting process (Pseudocode 7) is also undertaken in the watermark detector, we want to recovery the message sensed data from the cover medium. Based on the statistical rule of false-negative, we accept H_1 that means the cover medium is watermarked.

We then can do the process of extracting a watermark message sensed data into sensed data as shown in Figure 4. By using the pseudo code 7 the value of errors the cover medium with watermark constraints, we check whether these watermark constraints do not change or not. If these constraints do not change, we can do the process of extracting watermark signal. In this case, the coefficients objective function form the cover medium are 1 14 0 13 12 12 and 7.

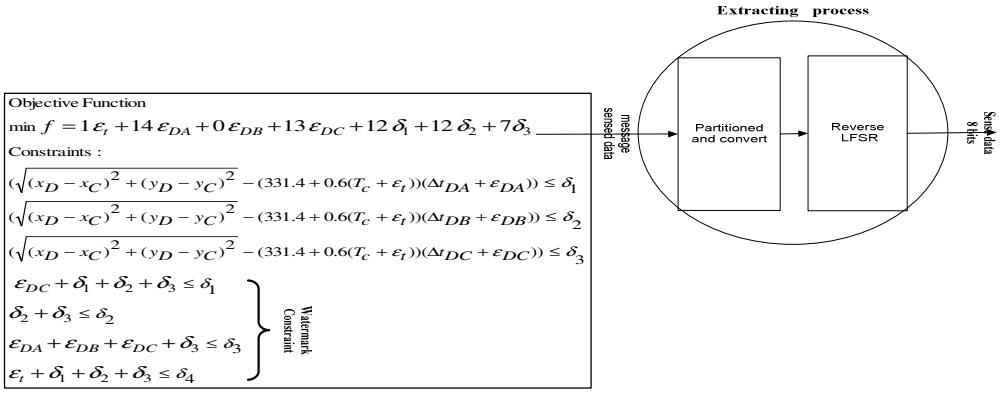


Figure 4 Watermark extracting Process

Pseudocode 7. The process of extracting sensed data

Input : $\varepsilon_t, \varepsilon_{DA}, \varepsilon_{DB}, \varepsilon_{DC}\ \delta_1, \delta_2, \delta_3$ watermark key
Output : Sensed data
Steps :
1. Compute the value of the objective f using $\varepsilon_t, \varepsilon_{DA}, \varepsilon_{DB}, \varepsilon_{DC}\ \delta_1, \delta_2$, and δ_3
2. If the value of the objective do not change go to 3
3. Else the value of the objective change goes to step 1.
4. Take the coefficients of objective f.
5. Convert the coefficient of objective f into 4 bits each.
6. Merge all of these 4 bits to 28 bits
7. Use reverse LFSR with watermark key to get sensed data.

6. EXPERIMENT SETUP

In this section, we describe the experiment setup for testing the purpose of the secure data transmission model, based on watermarking technique. We used TOMLAB which is a general purpose development environment in MATLAB for research and practical solution of optimization problems. TOMLAB has grown out of the need for advanced, robust and reliable

tools to be used in the development of algorithms and software for the solution of many different types of applied optimization problems.

6.1 NETWORK SETUP

In this section, the scenario of the atomic trilateration process is used as shown in Figure 4

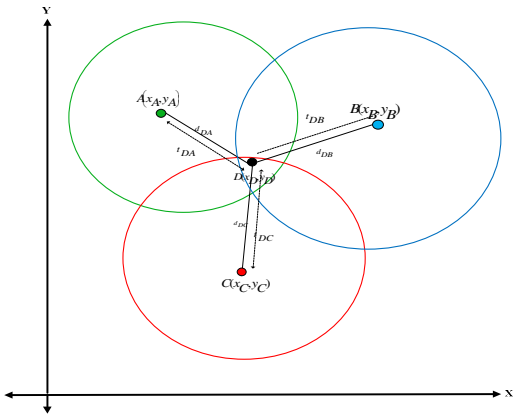


Figure 4 Atomic trilateration

With respect to a two-dimensional sensor networks, atomic trilateration is the means by which a sensor node in a networks can be used to determine its position by using the position of and distances to at least three other multimedia sensor nodes of know location. From these distance and position, a multimedia sensor node which is trying to determine its location can generate a nonlinear system equation. A typical scenario of atomic trilateration can be shown in Figure 4. Sensor node *D* trilaterates with another three sensor nodes *A*, *B*, and *C* which have coordinates $(x_A, y_A), (x_B, y_B)$, and (x_C, y_C). The distance is computed using time differences of arrival (TDoA) between acoustic signals simultaneously, which are emitted from a sensor nodes and received at the node *D* and radio frequency (RF). The sensor node *D* turns on a timer upon receiving the RF signal from the sensor node to measure the difference between the arrival of the RF and acoustic signals from that sensor node. The time measurements have an error. The speed of the acoustic signal is a function of the temperature of the propagation media. The relationship between the speed of the acoustic signal *Vs* (m/s) and the temperature T_c is as follows:

$$V_S = 331.4 + 0.6T_C \qquad \text{Eq.(4)}$$

By using the pseudocode 1, we find that the objective function is to minimize the overall error in the system, and can be stated as shown in Equation (1)

6.2 PERFORMANCE METRICS

The existing performance of the watermarking technique for secure data transmitting is evaluated against the following performance metrics:

Table 3 Performance Metrics secure data transmitting

Parameter	explain	Metric	Value
Node Sensor	Number of sensor node	Integer	100
(x_i, y_j), $i = j = 1,2,..n$	Position of two-dimensional sensor networks	Coordinate	x_i =115,5693 y_i =273,2856
T_c	the temperature of the propagation media	Degree	T_c =36
t_{DA}, t_{DB} t_{DC}	time transmission between node D to A, D to B and D to C	second	$t_{DA} = 0{,}771625$ $t_{DB} = 0{,}106793$ $t_{DC} = 0{,}09282$
Vs	Speed acoustic signal	(m/s)	$V_s \geq 331.4$
ε_t	the error in the measurement of the temperature	-	ε_t =0
ε_{DA}, ε_{DB}, ε_{DC}	the error in the measurement of the timer from D to A , D to B and D to C	-	$\varepsilon_{DA} = 0.0473$ ε_{DB} = -0.0141 , $\varepsilon_{DC} = 0$
δ_1 , δ_2 , δ_3	the error in the measurement between the Euclidean measurement and the measured using time differences of optimal D to A, D to B and D to C.	-	$\delta_1 = 0, \delta_2 = 0$ $\delta_3 = 0$
$\tau_1, \tau 2, \tau_3$ τ_4	the values are selected such that the feasibility of the solution space of the optimization problem is not harmed	-	τ_1 =0.16947616 τ_2 = 0.16947616, τ_3 = 0.24915965 τ_4 = 0.992920660
Sensed data	Data sensed by a sensor node	Bit	01111000
Watermark signal	Result from LFSR	Bit	00011110 000011011100 1100
Message sensed data	Result from pseudo code 4	Integer	1 14 0 13 12 12 7
treshold	normalized correlation the results of error the cover medium with watermark constraints	-	0.799153536405721
$sim(C', X')$	normalized correlation the results of error the cover medium with watermark constraints attack	-	0.2.154207742903002

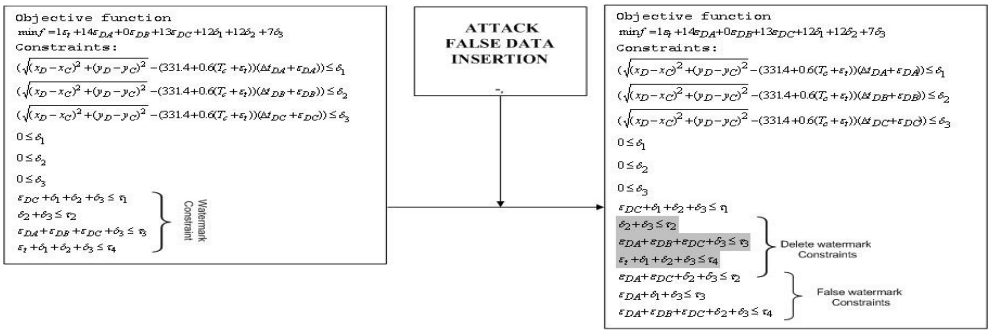
7. EXPERIMENT AND RESULTS FOR SECURE DATA TRANSMITTING

WSNs have an additionally vulnerability because node are often placed in a hostile environment where they are not physically protected. An attack is considered successful if it is not detected by the receiver. In this section we discuss various types of attacks that can be launched in the WSNs scenario and how the proposed security scheme can be used to thwart these attacks.

We consider in detail the corresponding weakness for this model watermarking technique that could be used by the attacker. Assume that the watermarks constraints are estimated by the attacker that should be change, modify and remove. The corresponding attacks are:

7.1 False data insertion attack

A number of different watermarks constraints that are generated by the LFSR, hoping to find the new results of error the cover medium that will map into existing solution.

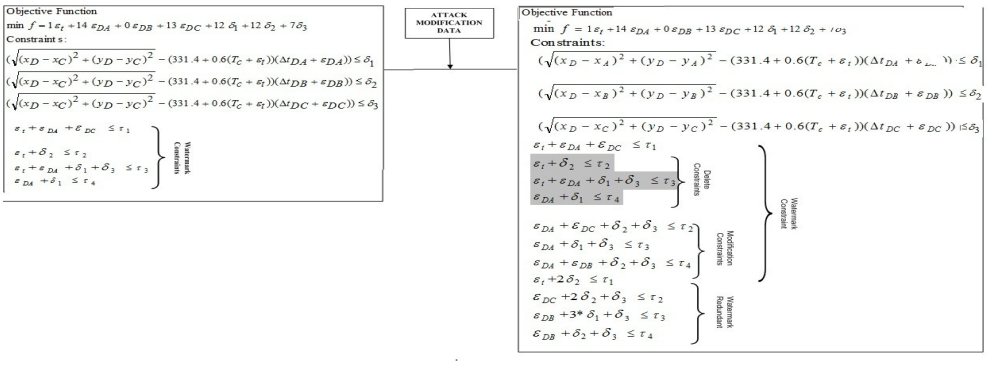


We get the results of the error of the cover medium by false insertion watermark constraints: ε_t =2.154233085444387, ε_{DA} = 0.007135532211399, ε_{DB} =- -0.000927368803587, ε_{DC} = 0.001724459319967, $\delta_1 = \delta_2 = 0$ and $\delta_3 = 0$.

Implementing a pseudo-code 6, we conclude that the value of similarity is greater than the value of threshold: the value of similarity = 2.154207742903002e+002>the value of threshold = 0.799153536405721. This means that the watermark signal is not robust to false data insertion attack.

7.2 Data modification attack

Data modification attack makes impersonation of different watermarks constraints that are generated by the LFSR, hoping also to find the new results of error the cover medium that will map into existing solution.

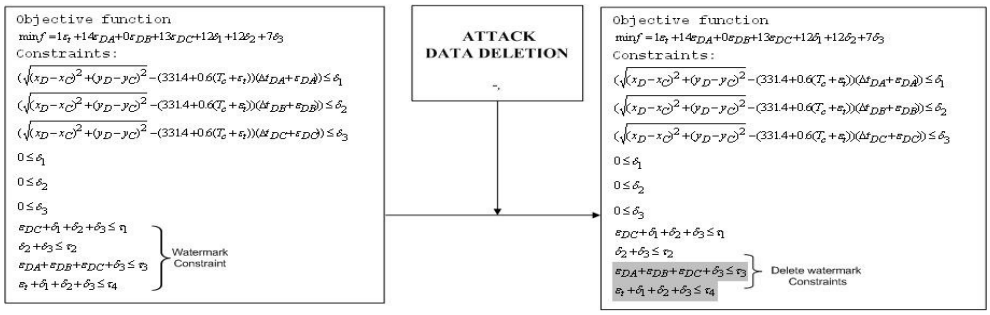


We get the results of the error of the cover medium by modification watermark constraints: ε_t =0.100170911928198, ε_{DA} = 1.118559233568045, ε_{DB} =- -0.167216683069220, ε_{DC} = 0.000000000000137, $\delta_1 = \delta_2 = 0$ and δ_3 = 0.145088180096586

Implementing a pseudo-code 6, we conclude that the value of similarity is greater than the value of threshold: the value of similarity =0.923139703988680 > the value of threshold = 0.799153536405721. This means that the watermark signal is not robust enough to modification the attack.

7.3 Data Deletion Attack

Data deletion attack is similar to the spoofed data attack in the sense that deleting watermark constraints make the error results of the cover medium invalid Delete a number of watermarks constraints hope to find new results of error the cover medium**.**

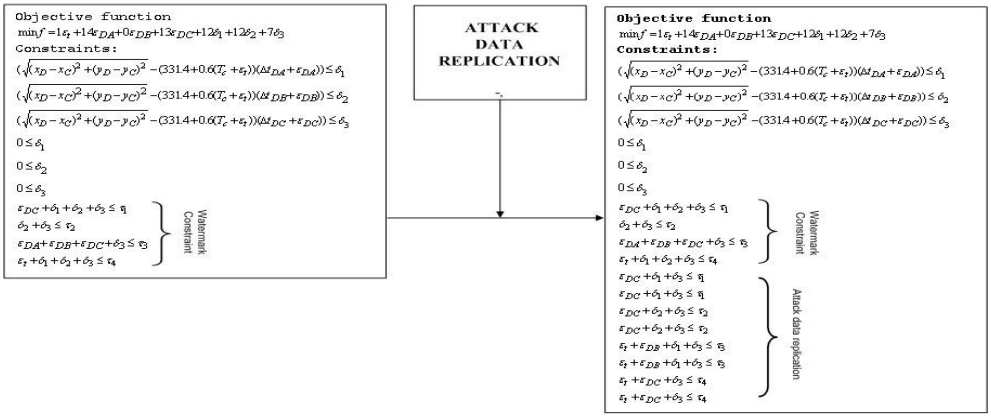


We get the results of the error of the cover medium by deleting watermark constraints: ε_t=0, ε_{DA} = 0.931857673282008, ε_{DB} =- 0.870667531648967, ε_{DC} = 0.145088180096586, $\delta_1 = \delta_2 = 0$ and $\delta_3 = 0$.

Implementing a pseudo-code 7, we conclude that the value of similarity is greater than the value of threshold: the value of similarity = 0.352844500181367< the value of threshold = 0.799153536405721. This means that the watermark signal is robust enough to delete the attack.

7.4 Replication Attack

Data replication attack is quite simple: an attacker seeks to add new constraints to the cover medium by replicating the new constraints with the existing constraints. New constraints replicated in this fashion can severely disrupt this solution of the cover medium's performance. Data replication attack hopes to find the new results of error the cover medium that will map into existing solution.

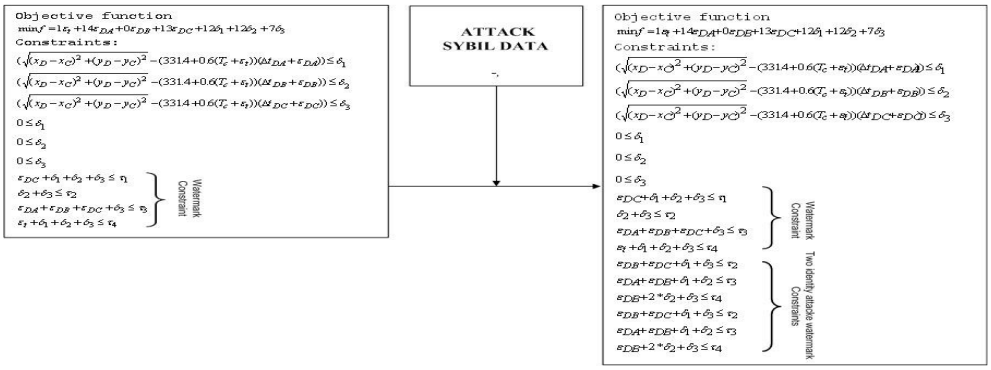


We get the results of the error of the cover medium by replication watermark constraints: ε_t=0.122299414900832, ε_{DA} = 0.428473573865247, ε_{DB} =- 0.500740473660944, ε_{DC} = 0, $\delta_1 = \delta_2 = 0$ and $\delta_3 = 0.145088180096586$.

Implementing a pseudo-code 7, we conclude that the value of similarity is greater than the value of threshold: the value of similarity = 0.285586856097203< the value of threshold = 0.799153536405721. This means that the watermark signal is robust to Replication Attack.

7.5 Sybil attack

A Sybil attack data occurs when the attacker creates multiple identities and exploits them in order to manipulate a reputation score. The Sybil attack data is defined as a malicious device illegitimately taking on multiple data identities., The Sybil attack data in communication channel watermarking is an attack wherein a reputation network system is subverted by forging more than one identity constraints in the cover medium A Sybil hopes to find aresults of error the cover medium.



We get the results of the error of the cover medium by Sybil watermark constraints: ε_I = 0.100170911928198, ε_{DA} = 0.118559233568045, ε_{DB} =- 0.013888456874134, ε_{DC} = 0.000000000000137, $\delta_1 = \delta_2 = 0$ and $\delta_3 = 0$.

Implementing a pseudo-code 7, we conclude that the value of similarity is greater than the value of threshold: the value of similarity = 0.103640805769825 < the value of threshold = 0.799153536405721. This means that the watermark signal is robust enough to Sybil attack the attack.

The results of these experiments have been shown in Table 3.

Table 3 The robustness of a watermark constraints, and watermark signal

No.	Kind of attacks	Watermark constraints	Watermark Signal
1.	False data insertion	Change	Not robust
2.	Data modification attack	Change	Not robust
3.	Data deletion	Not change	Robust
4.	Packet replication.	Not change	Robust
5.	Sybil attack	Not change	Robust

8. PERFORMANCE EVALUATION

In this section, we perform a comparative analysis of our technique with other techniques proposed by different researchers. The results of this comparative analysis are given in Table 4

Table 4 Comparative analysis with other approach

Kind of attacks	Feng et al. [5]	Sion et al. [6]	Koushan far et al. [8]	Zhang et al. [10]	Xiao et al. [9]	Xuejun et al. [11]	Kamel et al. [13]	Harjito B
False data insertion	x	x	x	√	x	x	√	x
data modification attack	x	x	x	√	x	x	√	x
Data deletion	x	x	x	x	x	x	√	√
Packet replication.	x	x	x	x	x	x	x	√
Sybil attack	x	x	x	x	x	x	x	√

√ provide secure data communication and robust x not provide secure data communication

We then do many experiments of these attacks above to test the performance of the model of secure data communication in WSNs. The results of these experiments can be shown in Table 3.

In this works, we compare 8 approaches in term of false data insertion, data modification attack, data deletion, packet replication, and Sybil attack. The [5], [6], [8], [9] approaches do not provide secure data communication against 6 attack. But [10] provide data for copy right protection and [13] provide for data integrity against these attacks. Our approach provides secure data communication against data deletion, packet replication and Sybil attacks. However our approach does not provide secure communication against false data insertion, and modification data

9. CONCLUSIONS

In this paper, we propose a watermarking technique for secure data transmitting in WSNs. Our strategy aims to protect data transmitting between sensor nodes in WSNs against these attacks. We verify our technique by brute force attacks. We can make secure data from data deletion, packet replication and Sybil attacks. However we cannot protect secure data from false data insertion, and modification data. Therefore, we still need to improve our technique under the circumstance that attacker launch different attack for the future work.

REFERENCES

[1] Yick, J., B. Mukherjee, and D. Ghosal, Wireless sensor network survey. Computer Networks, 2008. 52(12): p. 2292-2330.

[2] Jian, L. and H. Xiangjian. A Review Study on Digital Watermarking. in Information and Communication Technologies, 2005. ICICT 2005. First International Conference on. 2005.

[3] Rakesh, A. and K. Jerry, Watermarking relational databases, in Proceedings of the 28th international conference on Very Large Data Bases. 2002, VLDB Endowment: Hong Kong, China.

[4] Sion, R., M. Atallah, and P. Sunil, Rights protection for relational data. Knowledge and Data Engineering, IEEE Transactions on, 2004. 16(12): p. 1509-1525.

[5] Fang Jessica, P., Miodrag Real-time watermarking techniques for sensor networks Proceedings-SPIE The International Society for optical Engineering 2003(ISSU 5020): p. 391-402

[6] Radu, S., A. Mikhail, and P. Sunil, Resilient rights protection for sensor streams, in Proceedings of the Thirtieth international conference on Very large data bases - Volume 30. 2004, VLDB Endowment: Toronto, Canada.

[7] Xiao, X., Sun, Xingming Lincong, Yang Minggang, Chen. Secure Data Transmission of Wireless Sensor Network Based on Information Hiding. in Mobile and Ubiquitous Systems: Networking & Services, 2007. MobiQuitous 2007. Fourth Annual International Conference on. 2007.

[8] F. Koushanfar, M.P., Watermarking Technique for Sensor Networks: Foundations and Applications. Book chapter, in 'Security in Sensor Networks', Yang Xiao 2007.
[9] Rong, X., S. Xingming, and Y. Ying. Copyright Protection in Wireless Sensor Networks by Watermarking. in Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP '08 International Conference on. 2008.
[10] Zhang, W., Liu, Y, Sajal K, Das Aggregation Supportive Authentication in Wireless Sensor Networks: A Watermark Based Approach. in World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a. 2007.
[11] Xuejun, R., A sensitivity data communication protocol for WSN based on digital watermarking School of Information and Technology, Northwestern University, Xi'an 710127, China, 2010.
[12] Kamel, I.A.K., O. Al Dakkak , A. Distortion-Free Watermarking Scheme for Wireless Sensor Networks. in Intelligent Networking and Collaborative Systems, 2009. INCOS '09. International Conference on. 2009.
[13] Kamel, I., A Lightweight Data Integrity Scheme for Sensor Networks. Sensors, 2011. 11(4): p. 4118.
[14] Wang, X.-Y., Z.-H. Xu, and H.-Y. Yang, A robust image watermarking algorithm using SVR detection. Expert Systems with Applications, 2009. 36(5): p. 9056-9064.
[15] Potdar, V., Subjective and Objective Watermark Detection Using a Novel Approach–Barcode Watermarking ed. C.I.a. Security. 2007. 576.
[16] Vidyasagar, P., J. Christopher, and C. Elizabeth, Multiple image watermarking using the SILE approach, in Proceedings of the 6th WSEAS international conference on Multimedia systems \&\#38; signal processing. 2006, World Scientific and Engineering Academy and Society (WSEAS): Hangzhou, China.
[17] Albath, J., Practical algorithm for data security (PADS) in wireless sensor networks Proceedings of the 6th CM international workshop on Data engineering for wireless and mobile access - MobiDE '07. 2007. 9.
[18] Juma, H.K., I.Kaya, L. Watermarking sensor data for protecting the integrity. in Innovations in Information Technology, 2008. IIT 2008. International Conference on. 2008.
[19] Xiaomei, D.X., Li. An Authentication Method for Self Nodes Based on Watermarking in Wireless]Sensor Networks. in Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference on. 2009.
[20] Haowen, C. and A. Perrig, Security and privacy in sensor networks. Computer, 2003. 36(10): p. 103-105.
[21] Adrian, P.R., Szewczyk J. D. Tygar Victor, Wen David, E. Culler, SPINS: security protocols for sensor networks. Wirel. Netw., 2002. 8(5): p. 521-534.
[22] Bartosz, P., S. Dawn, and P. Adrian, SIA: secure information aggregation in sensor networks, in Proceedings of the 1st international conference on Embedded networked sensor systems. 2003, ACM: Los Angeles, California, USA.
[23] Xiaolong, L., et al., A Key Distribution Scheme Based on Public Key Cryptography for Sensor Networks, in Computational Intelligence and Security, W. Yuping, C. Yiu-Ming, and L. Hailin, Editors. 2007, Springer-Verlag. p. 725-732.
[24] Yao, J., A security architecture for wireless sensor networks based-on public key cryptography2009 5th International Conference on Wireless Communications, Networking and Mobile Computing. 2009. 1.
[25] Shih, F., Zhang, M, Towards Supporting Contextual Privacy in Body Sensor Networks for Health Monitoring Service. W3C Workshop on Privacy and data usage control, 04/05 October 2010 Cambridge (MA).
[26] E. Dawson, J.A., P. Gray, Australasian Journal of Combinatorics 1(1990), pp. 53-65.
[27] Harjito, B., Watermarking Technique based on Linear Feed Back Shift Register (LFSR), . Seminar Nasional Konferda ke –9 Himpunan Matematika Wilayah Jateng dan DIY di FMIPA UNS 2003.
[28] O'connor, J., and Roberstson, E,F Andrew nikolaevich Kolmogorof school of mathematics and Statistics, university of St Andrews, Scotland, 1999.
[29] Koushanfar, F. and M. Potkonjak, Watermarking Technique for Sensor Networks: Foundations and Applications. Book chapter, in 'Security in Sensor Networks', Yang Xiao (ed.), Auerbach publications 2006.
[30] Wong, J.L., Feng, J, Kirovski, D, Potkonjack M, Security in sensor networks: watermarking techniques in Wireless sensor networks. 2004. 305.
[31] Cox, I.J., Secure spread spectrum watermarking for multimedia. IEEE transactions on image processing, 1997. 6(12): p. 1673.

AUTHORS

Bambang Harjito is now as head of computer science department at Mathematics and Natural Science, Sebelas Maret University Surakarta, Indonesia. He received the master degree in computer science department at James Cook University in 2000 and He received PhD in School of information System, Curtin University Perth Australia in 2013.

Vidyasagar Potdar is a Senior Research Fellow working with School of Information Systems, Curtin Business School, Curtin University, Perth, Western Australia. He received the Bachelor of Science, Gujrat University, India 2001 and the Master of Informatoin Technology - University of Newcastle, Australia in 2002 and Doctor of Philosophy - Curtin University of Technology, Australia 2006. He is the Director of Anti-Spam Research Lab & Co-Director of Wireless Sensor Network Lab at the School of Information Systems.

OUTSOURCED KP-ABE WITH CHOSEN-CIPHERTEXT SECURITY

Chao Li, Bo Lang and Jinmiao Wang

State Key Laboratory of Software Development Environment,
Beihang University, Beijing, China
lichao@nlsde.buaa.edu.cn
langbo@buaa.edu.cn
wangjinmiao@nlsde.buaa.edu.cn

ABSTRACT

Key-Policy Attribute Based Encryption (KP-ABE) has always been criticized for its inefficiency drawbacks. Based on the cloud computing technology, computation outsourcing is one of the effective solution to this problem. Some papers have proposed their schemes; however, adversaries in their attack models were divided into two categories and they are assumed not to communicate with each other, which is obviously unrealistic. In this paper, we first proved there exist severe security vulnerabilities in these schemes for such an assumption, and then proposed a security enhanced Chosen Ciphertext Attack (SE-CCA) model, which eliminates the improper limitations. By utilizing Proxy Re-Encryption (PRE) and one-time signature technology, we also constructed a concrete KP-ABE outsourcing scheme (O-KP-ABE) and proved its security under SE-CCA model. Comparisons with existing schemes show that our constructions have obvious comprehensive advantages in security and efficiency.

KEYWORDS

KP-ABE, computation outsourcing, CCA, security model, proxy re-encryption, one-time signature

1. INTRODUCTION

With the development of Internet, the data storing service of third-party is increasingly popular. However, the data, in such a case, will be out of its owner's control and is managed by Cloud Storage Providers (CSPs). Then the confidentiality of the data becomes a problem. At present, encryption is the primary mechanism to implement data protection. However, traditional encryption schemes are not suitable in such a situation for their lack of ability in access control and huge overhead of key management. Sahai and Waters [1] addressed this issue by introducing the notion of attribute-based encryption (ABE), a new kind of public key based one-to-many encryption scheme which can achieve fine-grained access control on ciphertexts. In such a cryptosystem, private keys and ciphertexts are associated with an attribute group or an access policy respectively. A user is able to decrypt a ciphertext if and only if the attribute group satisfies the access policy. ABE can be classified as KP-ABE [2] and Ciphertext-Policy Attribute Based Encryption (CP-ABE) [3]-[5]. In KP-ABE, user's private key is associated with an access policy and ciphertexts are associated with a group of attributes. CP-ABE is just the opposite. They are suitable for different application scenarios. The former is data-centred with data attributes; the latter is user-centred with user attributes.

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 147–160, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41211

Although ABE is promising in implementing fine-grained access control on ciphertexts, it is being criticized for its inefficiency, which is first reflected in the process of decryption. The decryption of ABE is based on time-consuming bilinear pairings of which the number is in proportion to the complexity of the access policy. While conventional desktop computers would be able to handle such a task, it presents a significant challenge for users that manage and view private data on mobile devices. The inefficiency is also reflected in key-issuing. In applications that use ABE, the user groups are dynamically changed and the attribute universes of the user are usually very large. Hence, the heavy work of key generating during the initialization and user revocation will make Public Key Generator (PKG) a bo/ttleneck of the whole system. As the key is bundled with the policy in KP-ABE, more tasks will be needed in key generating a key and it will be more difficult to recognize the users affected by one revocation. Thus the inefficiency problem of key-issuing is more serious to KP-ABE.

The main solution to solve the problems above is outsourcing, by which we can outsource the heavy computation tasks to a third party who has strong computing power. And, the rise of cloud computing has provided techniques and application fundamentals for this. Cloud Service Providers (CSPs) could provide users the pay-on-demand computing services, such as Amazon's EC2 and Microsoft's Windows Azure. Based on this idea, Green et al. [6] firstly proposed the concrete ABE outsourcing schemes. In these schemes, all of the pairings in decryption are outsourced to a third party and cleartexts can be recovered by a simple ElGamal decryption without leaking any useful information of cleartexts and user private keys. We henceforth refer to this paper as Green11 [6].

However, Green11 cannot solve the inefficiency problem of key-issuing. Li et al. [7] extended the outsourcing idea to key-issuing of KP-ABE and proposed a new scheme model called Outsourced ABE (OABE). In OABE, there are three CSPs: S-CSP, D-CSP and KG-CSP, and they provide services of ciphertexts storing, decrypting and key generating respectively. We henceforth refer to this paper as LCLJ13. LCLJ13 can outsource both decryption and key-issuing. However, two pairings are still needed in the decryption phase. The authors of another paper [8] whose target is the checkability of outsourced results have improved the decryption efficiency by adopting the key-blinding technique. But the key-blinding work is done by PKG, thus it hasn't eased the PKG's burden compared to the traditional non-outsourcing KP-ABE. We refer to this paper as LHLC13.

Although the aforementioned schemes have alleviated the inefficiency problem of ABE to some degree, all of them fall short on the security, especially for LCLJ13 and LHLC13. Green11 can resist the Replayable Chosen Ciphertext Attack (RCCA) [9], which lies between Chosen Plaintext Attack and Chosen Ciphertext Attack. And we will explain RCCA in more detail in Section 6. LCLJ13 and LHLC13 share the same attack model, in which the adversaries are classified into two types: a curious user colluding with D-CSP and a curious KG-CSP. And they both assumed that the two types of adversaries cannot collude. However, in the real OABE system, all CSPs cannot be completely trusted. Thus, such an assumption is unrealistic. Actually, in their schemes, a curious user can decrypt any ciphertext by colluding with KG-CSP. We will give the proof in Section 4.1. Therefore, it is very important to construct a more secure KP-ABE outsourcing scheme which can outsource both key-issuing and decryption.

Our contributions. This paper focus on the KP-ABE outsourcing scheme which can outsource both key-issuing and decryption. Although LCLJ13 and LHLC13 can do this, they still have much space in improving efficiency and security, especially the security. We will firstly prove the security vulnerabilities in LCLJ13 and LHLC13 for their assumption of no collusion between curious users and KP-CSP. We further propose a security enhanced CCA (SE-CCA) model based on the analysis of environment with Chosen Ciphertext Attack, all CSPs are suspect and curious users may collude with any CSP. Then by utilizing the technique of PRE and one-time signature,

we construct a new concrete KP-ABE outsourcing scheme (O-KP-ABE) with proved security under SE-CCA.

Organization. The rest of the paper is organized as follows. In Section 2, we introduce the related work. Next, we give necessary background information in Section 3. In Section 4, we first give a detailed security analysis of existing schemes and then we describe our new KP-ABE outsourcing model and security enhanced CCA (SE-CCA) model. We present a concrete construction of a new KP-ABE outsourcing scheme (O-KP-ABE) and prove its security under SE-CCA in Section 5. In Section 6, we compare our scheme with all relative schemes in security and efficiency and analyse the results. Finally, we conclude our work.

2. RELATED WORK

ABE Outsourcing. This idea was firstly proposed by Green et al. [6] in their work. They have also constructed concrete schemes which can outsource the decryption based on this idea. Later, Zhou et al. [10] proposed a different ABE scheme with outsourced encryption and decryption. Zhou11 and Green11 both utilized the key-blinding technique to outsource decryption, in which the user firstly chooses a value randomly as the blind factor, and then runs exponentiations on the original key components with the blind factor. However, both of them haven't considered the computation overhead at PKG. Li et al. [7] firstly constructed a KP-ABE outsourcing scheme which can outsource both key-issuing and decryption. And the technique they adopted was different, of which the core idea is using a default attribute. This default attribute will be appended to each data's attribute group and each user's access policy. Besides, there are also papers [8], [11] researching on the verification of outsourcing results. And the main means is to append a redundancy to the ciphertext.

Proxy Re-Encryption (PRE). The notion of PRE was firstly proposed by Blaze et al. [12]. They also constructed a simple concrete scheme. PRE can be represented by the formula $D(\prod(E(m,e_A),\pi_{A\to B}),d_B)=m$, which means the ciphertext encrypted by A' s public key e_A after being re-encrypted by proxy key $\pi_{A\to B}$ can be decrypted by B' s secret key d_B . $\pi_{A\to B}$ is public and the re-encryption work can be done by an untrusted proxy server without fearing the leakage of the message *m*, and user secret keys d_A, d_B. Then Ateniese et al. [13] have made a further research on PRE, and they have concluded the features of a PRE scheme. Besides, they have also put forward an improved PRE scheme. There are also some papers [14] make research on the more secure PRE schemes.

One-Time Signature. There are several techniques to construct a CCA secure scheme from a CPA secure one. One-time signature is one of them. Canetti et al. [15] proposed this technique and utilized it to construct a CCA secure public key encryption scheme based on Identity-Based Encryption (IBE) [15]. Then Cheung and Newport [16] applied the similar technique to CP-ABE and constructed a CCA secure CP-ABE scheme from the CPA secure one. One-time signature contains a pair of keys (*sk, vk*) in which the former is used for signing and the latter is used for verifying and its length in binary bits is constant. In the scheme of Cheung et al., every bit in *vk* is defined as an attribute, thus there are two attributes corresponding to each bit. Each user secret key contains two components for both occurrences of each bit except for ABE components. For encryption, the encryptor chooses a pair (*sk, vk*) and encrypts the message with *vk* in addition to other attributes. The whole ciphertext is then signed with *sk*. And the decryptor will first verify the signature before decryption. We will take advantage of this technique to construct the scheme of O-KP-ABE and prove its CCA security under SE-CCA model.

3. PRELIMINARIES

3.1. Bilinear Maps

Let G and G_T be two multiplicative cyclic groups of prime order p and g is a generator of G. $e: G \times G \to G_T$ is a bilinear map with the properties:

- Bilinearity: for all $u, v \in G$ and $a, b \in Z_p$, we have $e(u^a, v^b) = e(u, v)^{ab}$.
- Non-degeneracy: $e(g, g) \neq 1$.

We say that G is a bilinear group if the group operation in G and the bilinear map $e: G \times G \to G_T$ are both efficiently computable. Notice that the map e is symmetric since $e(u^a, v^b) = e(u, v)^{ab} = e(u^b, v^a)$.

3.2. DBDH Assumption

We define the decisional Bilinear Diffie-Hellman problem as follows. A challenger chooses a group G of prime order p according to the security parameter. Let $a, b, c \in Z_p$ be chosen at random and g be a generator of G. Given (g, g^a, g^b, g^c), the adversary must distinguish a valid tuple $e(g, g)^{abc} \in G_T$ from a random element R in G_T.

Then we can get the definition of DBDH assumption:

Definition 1 (DBDH Assumption) *We say that the DBDH assumption holds if no polytime algorithm has a non-negligible advantage in solving the DBDH problem.*

3.3. Access Structure

Definition 2 (Access Structure [17]) *Let* $\{P_1, P_2, ..., P_n\}$ *be a set of parties. A collection* $A \subseteq 2^{\{P_1, P_2, ..., P_n\}}$ *is monotone if* $\forall B, C$ *if* $B \in A$ *and* $B \subseteq C$ *then* $C \in A$. *An access structure (respectively, monotone access structure) is a collection (respectively, monotone collection)* A *of non-empty subsets of* $\{P_1, P_2, ..., P_n\}$, *i.e.,* $A \subseteq 2^{\{P_1, P_2, ..., P_n\}} \setminus \varnothing$. *The sets in* A *are called the authorized sets, and the sets not in* A *are called the unauthorized sets.*

In the context of ABE, the role of the parties is taken by the attributes. Thus, the access structure A will contain the authorized sets of attributes.

4. NEW MODELS FOR KP-ABE WITH OUTSOURCING

4.1. Security Analysis of Existing Schemes

We will take LCLJ13 for example to prove it in this section, that is, if curious users collude with KG-CSP, they can decrypt any ciphertext CT. LHLC13 suffers from the same problem. The full proof is shown as follows.

Assume a curious user will collude with KG-CSP, and his secret key is SK = (SK_1, SK_2), in which $SK_2 = (d_{\theta 0} = g_2^{x_2} (g_1 h)^{r_\theta}, d_{\theta 1} = g^{r_\theta})$. As KG-CSP computes all delegated key-issuing work,

it may store the copies of all OKs, including the curious user's, and assume it is OK = x_1. Given a ciphertext CT= $(C_0 = m.e(g_1, g_2)^s, C_1 = g^s, E_\theta = (g_1 h)^s)$, the curious user performs the following steps:

(1) With public parameter g_2, OK and C_1 in CT, he can calculate $e(g, g_2)^{x_1 s}$; with SK2 and E_θ in CT, he can calculate $e(C_1, d_{\theta 0}) / e(d_{\theta 1}, E_\theta) = e(g, g_2)^{x_2 s}$.

(2) As the master key is $x = x_1 + x_2$, he is able to get $e(g, g_2)^{xs}$ through calculation.

(3) As $C_0 = m.e(g, g_2)^{xs}$, the curious user can recover *m*.

As CT in the above process can be any ciphertext, the curious user can decrypt all ciphertexts by colluding with KG-CSP. This is obviously incorrect. Thus there exist severe security vulnerabilities in LCLJ13 and LHLC13.

4.2. Model of KP-ABE with Outsourcing

In this section, we give our KP-ABE outsourcing model by modifying the model of KP-ABE with outsourced decryption in Green11. Our model supports the outsourcing of decryption and key-issuing simultaneously. The model is similar to LCLJ13, but it needs no subsequent processing of outsourced key-issuing, i.e. the PKG in our model need not do any further computations after receiving TK from KG-CSP. By distinguishing OK and TK, on one hand the user could decrypt the ciphertext himself when the network is unavailable; on the other hand the D-CSP need not to generate TK from OK whenever it translates ciphertexts. Our new KP-ABE outsourcing scheme consists of six algorithms, rather than seven in LCLJ13.

Setup. The setup algorithm takes no input other than the implicit security parameter. It is used to initialize the system and output the public parameter PK and master key MK. This algorithm is performed by PKG.

Encrypt (PK, M, S). The encryption algorithm takes as input the public parameters PK, a message M and a set of attributes S. It will encrypt M and produce a ciphertext CT. This algorithm is performed by Data Owner (DO).

Keygen_IN (A , MK, PK). This algorithm is the first step of key-issuing and is performed by PKG. It takes as input the access structure A , the master key MK and the public parameters PK. It outputs the outsourcing key OK and user private key SK.

Keygen_OUT (OK, PK). This algorithm is the second step of key-issuing and is performed by KG-CSP. It takes as input the outsourcing key OK and public parameters PK. It will output the transformation key TK and return it to PKG.

Transform_OUT (TK, CT). This algorithm completes the preprocessing of ciphertext and is performed by D-CSP. It firstly checks whether the attribute set S in CT satisfies the access structure A in TK. It outputs the partially decrypted ciphertext CT' if $S \in A$ otherwise it outputs $\perp$.

Decrypt (CT, CT', SK). This algorithm takes the ciphertext CT, partially decrypted ciphertext CT', and user private key SK as input. It outputs the message M if $S \in A$, otherwise $\perp$.

4.3. Enhanced Security Model

This section analyzes all possible attacks to the KP-ABE outsourcing model given in section 4.2 under CCA and proposes a new Security Enhanced CCA model SE-CCA.

As the above outsourcing model outsources the majority of work during key-issuing and decrypting to a third party who is not completely trusted, more information may be leaked. Even though the computations of key-issuing and decryption are outsourced to different parties, they may collude with each other. Thus, the outsourcing model above will face attacks different from any previous ones, which results in a different attack model.

Through careful analysis of the new outsourcing model, we find attackers under CCA may get the following information or services:

- Like the basic ABE schemes, the attacker is able to achieve the service of key-issuing, and thus get the key pair (SK, TK) corresponding to the specific access policy A.
- Since KG-CSP is not trusted, it may save the copies of all OKs sent from PKG and the corresponding TKs. So the attacker may get all of the key pairs (OK, TK).
- Combining the above two points, the attacker can get the tuple of keys (OK, TK, SK) corresponding to the specific policy A.
- As S-CSP and D-CSP are both untrusted and the TK corresponding to any access policy A can be achieved, the attacker is able to get the transforming service to all ciphertexts.
- The adversary can get specific decrypting services under CCA model.

Based on these observations, we propose the new CCA model SE-CCA, and the model is defined as follows:

Init. The adversary A declares the set of attributes S^* and submits it to the challenger C.
Setup. The challenger C runs the **Setup** algorithm of KP-ABE outsourcing scheme and sends the public parameters PK to adversary A.

Phase 1. The adversary A is allowed to make any of the following queries repeatedly:

i. Query for (SK, OK, TK) corresponding to the access structure A with the restriction that for all $x \in Y_A$, $x \notin S^*$, in which Y_A is the collection of the attributes in A.
ii. Query for (OK, TK) corresponding to the access structure A, with the restriction that for all $x \in Y_A$, $x \in S^*$, in which Y_A is the collection of the attributes in A.
iii. Query for the transforming ciphertext CT' corresponding to CT encrypted with S^*.
iv. A submits a ciphertext CT encrypted with S^* and gets the corresponding message m.

Challenge. A sends to C two equal length messages m_0, m_1. Then C flips a random coin b, and encrypts m_b with S^*. The ciphertext CT^* will be sent to A.

Phase 2. The adversary repeats **Phase 1** with the restriction that the ciphertexts submitted for decryption are not equal to CT^*.
Guess. A outputs a guess b' of b.

Definition 3 (SE-CCA Secure KP-ABE with Outsourcing) *An KP-ABE outsourcing scheme is SE-CCA secure if all polynomial time adversaries have at most a negligible advantage in the game of SE-CCA.*

5. O-KP-ABE

5.1. Access Trees

Our construction uses the tree-based access structure which is represented by *T*. Each interior node of the tree is a threshold gate and the leaves are associated with attributes. This structure is very expressive. For example, we can represent a tree with "AND" and "OR" gates by using respectively 2 of 2 and 1 of 2 threshold gates. A user is able to decrypt a ciphertext if and only if the attributes in ciphertext satisfies the access structure in the user's private key. The definitions of *T* and the relative functions are identical to paper [2].

5.2. Construction of O-KP-ABE

Let G_1 be a bilinear group of prime order p and let g be a generator of G_1. In addition, let $e: G_1 \times G_1 \to G_T$ denote the bilinear map. We also define the Lagrange coefficient $D_{i,S}$ for $i \in Z_p$ and a subset, S, of Z_p: $D_{i,S}(x) = \prod_{j \in S, j \neq i} \frac{x-j}{i-j}$. Assume the length of *vk* in one-time signature is l and vk_j is the value of j_{th} bit in *vk*.

Our construction consists of 6 algorithms.

Setup. First, choose a bilinear group G_1 of prime order p with a generator g and a bilinear map $e: G_1 \times G_1 \to G_T$. Next, determine the universe of attributes according to the actual situation $U = \{a_1, a_2, ..., a_n\}$, and let i represents the index of attribute a_i in *U*. Then, choose $\alpha, t_i, \beta_i, \omega, u_j \in Z_p, 1 \le i \le n, 1 \le j \le 2l$ and $g_2 \in G_1$, in which t_i and β_i correspond to a_i. Then the master key is $MK = (\alpha, \omega, t_i, \beta_i, u_j), 1 \le i \le n, 1 \le j \le 2l$, and the public parameters PK are $PK = \{U, g, g_1 = g^{\alpha}, g_2, T_i = g^{t_i}, P_i = g_2^{t_i^{-1}\beta_i}, U_j = g^{u_j}\}, 1 \le i \le n, 1 \le j \le 2l$.

Encrypt (PK, M, S). Run the key generating algorithm of one-time-signature to get a pair of key (*sk, vk*) and randomly choose a value $s \in Z_p$. For j_{th} bit in *vk*, if $vk_j = 0$, calculate $E_j = U_j^s$; if $vk_j = 1$, calculate $E_j = U_{j+l}^s$. Thus, we can get $C = \{S, C_0 = M.e(g_1, g_2)^s, \{C_y = T_i^s\}_{y \in S}, E\}$, in which $E = \{E_j\}, 1 \le j \le l$, *y* represents an attribute and *i* is the index of *y* in the universe of attributes U.

Then sign on *C* with *sk* and obtain a signature σ. The final ciphertext is $CT = (C, \sigma, vk)$.

Keygen_IN (T, MK, PK). The algorithm proceeds as follows. First choose a random value $z \in Z_p$ and calculate $\delta = (\alpha - \omega)/z$. Then, choose a polynomial q_x for each node x (including the leaves) in the tree T. These polynomials are chosen in the following way in a top-down manner, starting from the root node *r*.

For each node x in the tree, set the degree η_x of the polynomial q_x to be one less than the threshold value k_x of that node, that is $\eta_x = k_x - 1$. Then, for the root node *r*, set $q_r(0) = \delta$ and η_r other points of the polynomial q_r randomly to define it completely. For any other node x, set $q_x(0) = q_{parent(x)}(index(x))$ and choose η_x other points randomly to completely define q_x.

Once the polynomials have been decided, for each leaf node *x*, we can get the value of $q_x(0)$, and then calculate $d_x = q_x(0) / \beta_i$, in which *i* is the index of *x* in the universe of attributes.

Then, randomly choose $\omega_j \in Z_p, 1 \le j \le l-1$, and calculate $\omega_l = \omega - \sum_{j=1}^{l-1} \omega_j$. Afterwards, calculate $\varphi_j = \omega_j / u_j$, $\xi_j = \omega_j / u_{j+l}$, $1 \le j \le l$. Thus, we have $G = (G_{j0} = g_2^{\varphi_j}, G_{j1} = g_2^{\xi_j})$.

Finally, user's private key is SK = *z* and outsourcing key is $OK = \{T, \{d_x\}_{x \in Y_T}, G\}$.

Keygen_OUT (OK, PK). For each element dx in OK calculate $D_x = P_i^{d_x} = g_2^{t_i^{-1} \cdot q_x(0)}$, in which *i* is the index of attribute *x* in the universe of attributes U. The transformation key is:

$TK = \{T, \{D_x\}_{x \in Y_T}, \mathrm{G}\}$, in which Y_T is the attributes set of leaves in T.

Transform_OUT (TK, CT). First verify the signature σ. On failure, return $\perp$. Otherwise, proceed the transformation procedure which is defined as a recursive algorithm TransformNode (*x*, TK, CT). This recursive algorithm outputs a group element of G_T or $\perp$.

If the node *x* is a leaf node then:

$$TransformNode(x, TK, CT) = \begin{cases} e(D_x, C_x) = e(g_2^{t_i^{-1} \cdot q_x(0)}, g^{t_i \cdot s}) \\ = e(g, g_2)^{q_x(0).s} \quad \text{if } x \in S \\ \\ \perp \quad \text{otherwise} \end{cases}$$

If *x* is not a leaf node, the algorithm TransformNode (*x*, TK, CT) proceeds as follows: for all nodes *z* that are children of *x*, it calls TransformNode (*z*, TK, CT) and stores the output as F_z. Let S_x be an arbitrary k_x sized set of child nodes *z* such that $F_z \neq \perp$. If no such set exists then the node was not satisfied and the function returns $\perp$.

Otherwise, we compute:

$$\begin{aligned} & TransformNode(x, TK, CT) \\ &= \prod_{z \in S_x} F_z^{\mathrm{D}_{i.S_x'}(0)}, \quad \text{in which } \begin{matrix} i = index(z) \\ S_x' = \{index(z) : z \in S_x\} \end{matrix} \\ &= \prod_{z \in S_x} (e(g, g_2)^{s.q_{z(0)}})^{\mathrm{D}_{i.S_x'}(0)} \\ &= \prod_{z \in S_x} (e(g, g_2)^{s.q_{parent(z)}(index(z))})^{\mathrm{D}_{i.S_x'}(0)} \\ &= \prod_{z \in S_x} e(g, g_2)^{s.q_x(i).\mathrm{D}_{i.S_x'}(0)} \\ &= e(g, g_2)^{s.q_x(0)} \end{aligned}$$

If CT cannot satisfy TK, the algorithm returns $\perp$, otherwise we can get $CT_0' = e(g, g_2)^{s\delta}$. Then, for j_{th} bit in *vk*, if $vk_j = 0$, calculate $e(E_j, G_{j0}) = e(g^{u_j s}, g_2^{\omega_j / u_j}) = e(g, g_2)^{\omega_j s}$; if $vk_j = 1$, calculate

$e(E_j, G_{j1}) = e(g^{u_{j+l}s}, g_2^{\omega_j/u_{j+l}}) = e(g, g_2)^{\omega_j s}$. Then we can get $W' = \prod_{j=1}^{l} e(g, g_2)^{\omega_j s} = e(g, g_2)^{\omega s}$. Thus, we can get the final result $CT' = \{CT_0', W'\}$.

Decrypt (CT, CT', SK). If the user has the privilege to access the data, then upon receiving CT' from D-CSP, the user completes the decryption and gets a message $M = C_0 / (CT_0'^{SK}, W')$.

5.3. Proof of Security under SE-CCA

We prove the following theorem:

THEOREM 1. *If an adversary can break the scheme of O-KP-ABE under the SE-CCA model, then a simulator can be constructed to solve the DBDH problem with a non-negligible advantage.*

PROOF: Suppose there exists a polynomial adversary A who can attack our scheme under the SE-CCA model with advantage ε and the probability to forge an legal signature is Pr[forge], then we can build a simulator S who can win DBDH problem with a non-negligible advantage $\frac{\varepsilon}{2} - \frac{1}{2}\Pr[\text{forge}]$. The process of simulation is as follows:

The challenger C generates the tuple (A, B, C, Z). Then, the simulator S chooses a signature key pair (*sk*, vk**). The length of *vk** is *l* and vk_j* is the value of j_{th} bit in *vk**. We assume the universe of attributes, U, is defined.

Init. A chooses the set of attributes *S** it wishes to be challenged upon and sents it to S.

Setup. The simulator S sets $g_1 = A = g^{\alpha}$ (thus, $a = \alpha$) and $g_2 = B$. Then choose a random value $\omega \in Z_p$. For each $a_i \in$ U, S chooses random values $t_i', \beta_i' \in Z_p$. If $a_i \in S^*$, the simulator sets $T_i = g^{t_i'}$ and $P_i = B^{t_i'^{-1}.\beta_i'} = g_2^{t_i'^{-1}.\beta_i'}$ (thus, $t_i = t_i', \beta_i = \beta_i'$); if $a_i \notin S^*$, S sets $T_i = A^{t_i'}.g^{-\omega t_i'} = g^{(\alpha-\omega)t_i'}$ and $P_i = g^{\beta_i'/t_i'} = g^{(\alpha-\omega)\beta_i'/(\alpha-\omega)t_i'}$, (thus, $t_i = (\alpha-\omega)t_i', \beta_i = (\alpha-\omega)\beta_i'$).

Next, randomly choose $u_j' \in Z_p, 1 \le j \le 2l$. If $vk_j^* = 0$, set $u_j = u_j'$ and $u_{j+l} = bu_{j+l}'$; otherwise, set $u_j = bu_j'$ and $u_{j+l} = u_{j+l}'$. Then calculate $U_j = g^{u_j}$.

So the public parameters are $PK = (U, g, g_1, g_2, T_i, P_i, U_j), 1 \le i \le n, 1 \le j \le 2l$, and they will be sent to A.

Phase 1. The adversary A is allowed to make any of the following four queries repeatedly:

i. A submits an access tree *T* with the restriction that for all $x \in Y_T$, $x \notin S^*$, in which Y_T is the attributes set of leaves in *T*. And S must construct the corresponding key tuple (SK, OK, TK). S firstly chooses a random value $z \in Z_p$ and sets SK = *z*.

Then, set $q_r(0) = 1/z$ and calculate the value of $q_x(0)$ of each leaf node *x* in tree *T* following the steps of **Keygen_IN**. Next, let $Q_x(0) = (\alpha-\omega).q_x(0)$, thus $Q_r(0) = (\alpha-\omega)/z$. Since the simulator sets $\beta_i = (\alpha-\omega)\beta_i'$ for all $a_i \notin S^*$, we can calculate $d_x = Q_x(0)/\beta_i = q_x(0)/\beta_i'$. Afterwards, randomly choose $\omega_j \in Z_p$, $1 \le j \le l-1$, and calculate $\omega_l = \omega - \sum_{j=1}^{l-1} \omega_j$. If $vk_j^* = 0$, calculate $G_{j0} = g_2^{\varphi_j/u_j} = B^{\varphi_j/u_j}$ and $G_{j1} = g_2^{\varphi_j/u_{j+l}} = g_2^{\varphi_j/bu_{j+l}'} = g^{\varphi_j/u_{j+l}'}$; otherwise,

$G_{j0} = g_2^{\varphi_j/u_j} = g^{\varphi_j/u_j'}$ and $G_{j1} = g_2^{\varphi_j/u_{j+l}} = g_2^{\varphi_j/u_{j+l}'} = B^{\varphi_j/u_{j+l}'}$. Thus, we get $G = \{G_{j0}, G_{j1}\}$. The outsourcing key is OK=$\{T, \{d_x\}_{x \in Y_T}, G\}$.

For each element d_x in OK calculate $D_x = P_i^{d_x}$, and the transformation key is TK=$\{T, \{D_x\}_{x \in Y_T}, G\}$.

Finally, send (SK, OK, TK) to the adversary A.

ii. A submits an access tree *T* with the restriction that for all $x \in Y_T$, $x \in S^*$, in which Y_T is the attributes set of leaves in *T*. And S must construct the corresponding key tuple (OK, TK).

S firstly chooses a random value $z' \in Z_p$, and sets $\delta = 1/z'$ (thus $z = (\alpha - \omega).z'$). Then, set $q_r(0) = 1/z'$ and calculate the value of $q_x(0)$ of each leaf node *x* in tree *T* following the steps of **Keygen_IN**. Since the simulator sets $\beta_i = \beta_i'$ for all $x \in S^*$, we can calculate $d_x = q_x(0)/\beta_i = q_x(0)/\beta_i'$. Afterwards, generate *G* using the same approach in i. Thus, the outsourcing key is OK=$\{T, \{d_x\}_{x \in Y_T}, G\}$.

For each element d_x in OK calculate $D_x = P_i^{d_x}$, and the transformation key is TK=$\{T, \{D_x\}_{x \in Y_T}, G\}$.

iii. A submits a ciphertext CT encrypted by *S**, the simulator must transform it to CT'.

First, S verifies the correctness of σ. On failure, the simulator will terminate the game and return $\perp$, which is called the "Exist Event". Otherwise, generate an access tree and revoke ii to get the corresponding TK. Then with TK, S can transform CT to CT'.

iv. A submits a ciphertext CT=$\{C, \sigma, vk\}$ encrypted by *S**, and the simulator will decrypt it.

First, revoke query iii, if the result is not $\perp$, then proceed as follows:

- ✧ If *vk* = *vk**, we call the "Forgery" happens and the simulator will terminate the game and output the guess *u*' of *u* randomly.
- ✧ If $vk \neq vk^*$, we can assume the j_{th} bit of them are different. Without loss of generality, we assume $vk_j = 1$ and thus $vk_j^* = 0$. Therefore, $E_j = g^{u_{j+l}s} = g^{bu_{j+l}s}$. Then the simulator can calculate $e(A, E_j) = e(g^a, g^{bu_{j+l}'s} = e(g,g)^{abs.u_{j+l}'}$. Since u_{j+l}' is known to S, he calculates $C_0 / e(A, E_j)^{1/u_{j+l}'s}$ as the message m and sends it to the adversary.

Challenge. The adversary A submits two challenge messages m_0, m_1 with equal length to S. The simulator S will flip a fair binary coin *b*, and returns an encryption of m_b. The ciphertext is outputted as $C^* = \{S^*, C_0 = m_b.z, \{C_y = C^{t_i'}\}_{y \in S^*}, E\}$. E is calculated as follows: if $vk_j^* = 0$, $E_j = U_j^s = g^{u_j's} = C^{u_j'}$; otherwise, $E_j = U_{j+l}^s = g^{u_{j+l}'s} = C^{u_{j+l}'}$.

If $u = 0$ then $Z = e(g,g)^{abc}$. If we let $s = c$, then we have $C_0 = m_b.e(g,g)^{abc} = m_b.e(g^a, g^b)^c = m_b.e(g_1, g_2)^s$, $C_y = C^{t_i'} = g^{t_i's} = T_i^s$. Therefore, the ciphertext is a valid random encryption of message m_b.

If $u = 1$, then $Z = e(g,g)^z$. Thus, $C_0 = m_b.e(g,g)^z$. Since *z* is random, C_0 will be a random element of G_T from adversary's view and the message contains no information about m_b.

Then, sign *C** with *sk** to get the signature σ. The final ciphertext is $CT^* = \{C^*, \sigma, vk^*\}$ and it will be sent to the adversary.

Phase 2. Repeat the process of **Phase 1** with the restriction that the ciphertexts submitted for decryption are not equal to *CT**.

Guess. A will submit a guess *b'* of *b*. If *b'*=*b*, the simulator will output *u'*=0 to indicate it was given a valid BDH-tuple, otherwise it will output *u'*=1 to indicate it was given a random 4-tuple.

First of all, "forge" represents that the event of forgery happens and "¬forge" represents the opposite.

If "Forgery" happens, S will not wait for A 's guess of *b* and output the guess of *u* randomly, thus $\Pr(u' = u \mid \text{forge}) = \frac{1}{2}$.

If *u*=1 and "Forgery" doesn't happen, the adversary gains no information about *b*. Therefore, we have $\Pr(b \neq b' \mid \neg\text{forge} \mid u = 1) = \frac{1}{2}$. Since the simulator guess *u'*=1 when *u'*=*u*, we have $\Pr(u' = u \mid \neg\text{forge} \mid u = 1) = \frac{1}{2}$. Thus, the probability to solve DBDH problem for S when *u*=1 is:

$$\begin{aligned}
&\Pr(u' = u \mid u = 1) \\
&= \Pr(u' = u, \text{forge} \mid u = 1) + \Pr(u' = u, \neg\text{forge} \mid u = 1) \\
&= \Pr(u' = u \mid \text{forge} \mid u = 1).\Pr(\text{forge}) + \\
&\quad \Pr(u' = u \mid \neg\text{forge} \mid u = 1).\Pr(\neg\text{forge}) \\
&= \tfrac{1}{2}.\Pr(\text{forge}) + \tfrac{1}{2}.(1 - \Pr(\text{forge})) \\
&= \tfrac{1}{2}
\end{aligned}$$

If *u*=0, the adversary sees an valid encryption of m_b. The adversary's advantage in this situation is ε by definition. Therefore, we have $\Pr(b' = b \mid \neg\text{forge} \mid u = 0) = \frac{1}{2} + \varepsilon$. Since the simulator guess *u'*=0 when *b'*=*b*, we have $\Pr(u' = u \mid \neg\text{forge} \mid u = 0) = \frac{1}{2} + \varepsilon$. Thus, the probability to solve DBDH problem for S when *u*=0 is:

$$\begin{aligned}
&\Pr(u' = u \mid u = 0) \\
&= \Pr(u' = u, \text{forge} \mid u = 0) + \Pr(u' = u, \neg\text{forge} \mid u = 0) \\
&= \Pr(u' = u \mid \text{forge} \mid u = 0).\Pr(\text{forge}) + \\
&\quad \Pr(u' = u \mid \neg\text{forge} \mid u = 0).\Pr(\neg\text{forge}) \\
&= \tfrac{1}{2}.\Pr(\text{forge}) + (\tfrac{1}{2} + \varepsilon).(1 - \Pr(\text{forge})) \\
&= \tfrac{1}{2} + \varepsilon - \varepsilon.\Pr(\text{forge})
\end{aligned}$$

Thus, the overall advantage of simulator in the DBDH game is:

$$\begin{aligned}
&\tfrac{1}{2}\Pr(u' = u \mid u = 0) + \tfrac{1}{2}\Pr(u' = u \mid u = 1) - \tfrac{1}{2} \\
&= \tfrac{1}{2}.(\tfrac{1}{2} + \varepsilon - \varepsilon.\Pr(\text{forge})) + \tfrac{1}{2}.\tfrac{1}{2} - \tfrac{1}{2} \\
&= \tfrac{1}{2}.\varepsilon - \tfrac{1}{2}.\varepsilon.\Pr(\text{forge}) \\
&\geq \tfrac{1}{2}.\varepsilon - \tfrac{1}{2}.\Pr(\text{forge})
\end{aligned}$$

6. ANALYSIS AND DISCUSSION

6.1. Analysis

This section compares our scheme with other existing KP-ABE outsourcing schemes in efficiency and security. The results are shown in Table 1.

Table 1. Comparisons in efficiency and security between our schemes and others.

Scheme	KG Ops	Dec Ops	Security Level
Green11	SS+5|Y|G	1G	RCCA
LCLJ13	3G	2P	CPA
LHLC13	(2|Y|+5)G	1G	CPA
O-KP-ABE(ours)	SS+2*l*G	1G	SE-CCA

G and P stand for the maximum time to compute an exponentiation in G and a pairing respectively. |Y| denotes the number of leaves in access tree. l is the length of vk. SS represents the time to share a secret in key-issuing phase.

As Green11 has not considered the outsourcing of key-issuing, the number of exponentiations that it must accomplish is proportional to the size of the access tree. Thus, its efficiency of key-issuing is relatively low. LCLJ13 can outsource both decrypting and key-issuing and PKG only needs to complete three exponentiations during the key-issuing phase. However, the user still has to complete two pairings when decrypting ciphertexts. LHLC13 improves that, and the user only needs to complete one exponentiation in decryption. But its efficiency in key-issuing decreases sharply, even no better than the original scheme [2] without outsourcing.

Our O-KP-ABE scheme has the highest efficiency in decryption, where the user only needs one exponentiation. We have utilized the technique of one-time signature, and PKG must do extra 2l exponentiations. Thus, its efficiency in key-issuing is relatively low. That is the cost of security.

In the aspect of security, Green11 can resist the Replayable Chosen Ciphertext Attack (RCCA) [9]. The traditional notion of security against CCA is a bit too strong, since it does not allow any bit of the ciphertext to be altered. However, there exist encryption schemes that are not CCA secure, but seem sufficiently secure “for most practical purposes”. For these reasons, Canetti et al. proposed the notion of RCCA. On one hand, RCCA security accepts as more secure than some non-CCA schemes; on the other hand, it suffices for most of existing applications of CCA security. Thus, the security of RCCA lies between CPA and CCA. Although the authors of LCLJ13 and LHLC13 declared that their schemes are CPA secure, we have proved that they have severe security vulnerability when collusion is considered. Our scheme has been proved secure under the SE-CCA model, which means O-KP-ABE has removed the security vulnerability in LCLJ13 and LHLC13 and can be CCA secure. Thus our scheme has a relatively higher security compared with the existing schemes.

6.2. Discussions

Verifiability. Although in our scheme the proxy servers, namely CSPs, cannot learn anything useful, there is no guarantee on the correctness of the outsourcing results. In some applications users or PKGs often request to check whether the outsourcing work is indeed done correctly. This is another important issue in outsourcing KP-ABE, and some approaches have been proposed. For example, Lai et al. [11] and Li et al. [8] addressed this problem by appending a redundancy to the ciphertext. However, both of them only considered the verification of outsourced decrypting, they

have not considered the same request for outsourced key-issuing. In our future work, we will consider the verification issue of our O-KP-ABE scheme.

Similar Problems for CP-ABE. Despite the work like this paper that focus on the KP-ABE outsourcing scheme, there are also many papers [10] engaged in the outsourcing of encryption and decryption of CP-ABE. However, we find that some of them have the similar security vulnerabilities to LCLJ13 and LHLC13. For example, in Zhou et al.'s scheme [10], the ciphertext embedded with the access policy $T_{ESP} \wedge T_{DO}$ can be decrypted by any user that satisfies T_{DO} by colluding with ESP. Firstly, the user gets s_1 from ESP. Then he chooses a pair of key components $d = (D_j, D_j')$ from his secret key, and we assume that the corresponding attribute is y. After that, the user can compute the ciphertext components pair corresponding to y as $c = (C_y = g^{s_1}, C_y' = H(y)^{s_1})$. And then he can calculate $e(g,g)^{rs_1}$ with d and c. In addition, because the user satisfies T_{DO}, he can compute $e(g,g)^{rs_2}$. At last, with the public parameter h and the key component D, the user can restore the message in the ciphertext. Thus we can see that this scheme has severe security problems. However, if we replace the hash function H with the attributes universe, then the similar technique in our scheme can be used to solve this problem. We leave this problem to our future work.

7. CONCLUSION

It is unrealistic for the existing KP-ABE outsourcing schemes to assume that curious users will not collude with KG-CSP. Thus, in this paper we first proposed a new security enhanced model SE-CCA. All CSPs in SE-CCA are curious and allowed to collude with each other. Besides, the attackers can get decryption service for specific ciphertexts in SE-CCA. Then, we constructed a concrete outsourcing scheme O-KP-ABE and proved its security under SE-CCA. Our scheme has the highest security compared with existing schemes. Except for that, O-KP-ABE also has relative higher efficiency. Hence, our construction has a comprehensive advantage over existing schemes in security and efficiency.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No.61170088) and Foundation of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2013ZX-05).

REFERENCES

[1] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in Advances in Cryptology-EUROCRYPT 2005. Springer, 2005, pp. 457–473.

[2] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in Proceedings of the 13th ACM conference on Computer and communications security. ACM, 2006, pp. 89–98.

[3] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in Security and Privacy, 2007. SP'07. IEEE Symposium on. IEEE, 2007, pp. 321–334.

[4] L. Ibraimi, Q. Tang, P. Hartel, and W. Jonker, "Efficient and provable secure ciphertext-policy attribute-based encryption schemes," in Information Security Practice and Experience. Springer, 2009, pp. 1–12.

[5] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in Public Key Cryptography–PKC 2011.Springer, 2011, pp. 53–70.

[6] M. Green, S. Hohenberger, and B. Waters, "Outsourcing the decryption of abe ciphertexts." in USENIX Security Symposium, 2011, p. 3.

[7] J. Li, X. Chen, J. Li, C. Jia, J. Ma, and W. Lou, "Fine-grained access control system based on outsourced attribute-based encryption," in Computer Security–ESORICS 2013. Springer, 2013, pp. 592–609.

[8] J. Li, X. Huang, J. Li, X. Chen, and Y. Xiang, "Securely outsourcing attribute-based encryption with checkability," IEEE Transactions on Parallel and Distributed Systems, pp. 2201–2210, 2013.

[9] R. Canetti, H. Krawczyk, and J. B. Nielsen, "Relaxing chosen-ciphertext security," in Advances in Cryptology-CRYPTO 2003. Springer, 2003, pp. 565–582.

[10] Z. Zhou and D. Huang, "Efficient and secure data storage operations for mobile cloud computing," in Proceedings of the 8th International Conference on Network and Service Management. International Federation for Information Processing, 2012, pp. 37–45.

[11] J. Lai, R. H. Deng, C. Guan, and J. Weng, "Attribute-based encryption with verifiable outsourced decryption," Information Forensics and Security, IEEE Transactions on, vol. 8, no. 8, pp. 1343–1354, 2013.

[12] M. Blaze, G. Bleumer, and M. Strauss, "Divertible protocols and atomic proxy cryptography," in Advances in CryptologyEUROCRYPT'98. Springer, 1998, pp. 127–144.

[13] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage," ACM Transactions on Information and System Security (TISSEC), vol. 9, no. 1, pp. 1–30, 2006.

[14] R. Canetti and S. Hohenberger, "Chosen-ciphertext secure proxy re-encryption," in Proceedings of the 14th ACM conference on Computer and communications security. ACM, 2007, pp. 185–194.

[15] R. Canetti, S. Halevi, and J. Katz, "Chosen-ciphertext security from identity-based encryption," in Advances in Cryptology-Eurocrypt 2004. Springer, 2004, pp. 207–222.

[16] L. Cheung and C. Newport, "Provably secure ciphertext policy abe," in Proceedings of the 14th ACM conference on Computer and communications security. ACM, 2007, pp. 456–465.

[17] A. Beimel, "Secure schemes for secret sharing and key distribution," Ph.D. dissertation, Technion-Israel Institute of technology, Faculty of computer science, 1996.

LIVER SEGMENTATION FROM CT IMAGES USING A MODIFIED DISTANCE REGULARIZED LEVEL SET MODEL BASED ON A NOVEL BALLOON FORCE

Nuseiba M. Altarawneh[1],SuhuaiLuo[1], Brian Regan[1], Changming Sun[2]

[1] School of Design Communication and IT,
The University of Newcastle, Callaghan NSW 2308, Australia
[2]Computational Informatics, CSIRO, North Ryde, NSW 1670, Australia
nuseiba.altarawneh@uon.edu.au, suhuai.luo@newcastle.edu.au,
brian.regan@newcastle.edu.au, changming.sun@csiro.au

ABSTRACT

Organ segmentation from medical images is still an open problem and liver segmentation is a much more challenging task among other organ segmentations. This paper presents a liver segmentation method from a sequence of computer to mography images.We propose a novel balloon force that controls the direction of the evolution process and slows down the evolving contour in regions with weak or without edges and discourages the evolving contour from going far away from the liver boundary or from leaking at a region that has a weak edge, or does not have an edge. The model is implemented using a modified Distance Regularized Level Set (DRLS) model. The experimental results show that the method can achieve a satisfactory result. Comparing with the original DRLS model, our model is more effective in dealing with over segmentation problems.

KEYWORDS

Liver segmentation, level set method, Distance Regularized Level Set (DRLS) model

1. INTRODUCTION

The liver is one of the most important organs in the human body. It carries out a variety of functions including filtering the blood, making bile and proteins, processing sugar, breaking down medications, and storing iron, minerals and vitamins. However, the liver is prone to many diseases such as hepatitis C, cirrhosis, and cancer. As the advance of computer science and technology, computer-aided surgical planning systems (CAD) have played an important role in diagnosing and treatment of liver diseases. These systems can present the structures of various liver vessels, generate resection proposals, offer 3D visualizations, provide surgical simulations with cutting, and lead to shorter planning times. However, among these systems, one of the most

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 161–170, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41212

important problems is the accurate segmentation of a liver from its surrounding organs in computer tomography (CT) images.

Developing a robust method for liver segmentation from CT images is a challenging task due to the similar intensity values between adjacent organs, geometrically complex liver structure and the injection of contrast media, which causes all tissues to have different gray level values. Several artefacts of pulsation and motion, and partial volume effects also increase the difficulties to carry out automatic liver segmentation in CT images. The significant variations in shape and volume of the liver also contribute these difficulties. Therefore, liver segmentation from medical images is still an open problem. Generally, methods and approaches to liver segmentation in the CT images are categorized into two main categories: semiautomatic and fully automatic liver segmentation methods. Semi-automatic liver segmentation methods require a limited user intervention to complete the task. This intervention varies from a manual selection for seed points to a manual refinement of a binary mask for the liver. The term fully automated means that the liver segmentation process is implemented without any sort of operator intervention. This kind of method is highly appreciated by radiologists since it is free from user errors and biases, and it saves the operator from a potentially hard work and wasted time. The latest achievements in liver segmentation are reviewed in this section. All the methods are discussed in one of the three categories; namely, gray level based methods, model based methods, and texture based methods [1]. It is apparent that each category of liver segmentation methods has its own advantages and shortcomings, and may be effective for a particular case. With the gray level based methods[2-11], they use image feature directly. Consequently, gray level based methods are the most common methods in liver segmentation.

These methods depend primarily on the evolution of gray level toward targets. While gray level methods are generally fast, their effectiveness may be affected, especially when gray level intensity of targets exhibits changes. In spite of using prior knowledge, gray level methods may fail when the liver occupies a small percentage of the image. Gray level is applied manually or via automatic rough segmentation. The aim of these two procedures is to collect information with regard to the gray level. Whilst these methods are reliable, they often require substantial computational time. Several gray level based methods deploy gradient information as a precise approach to deal with image boundaries. However, this approach becomes impractical in the presence of numerous boundaries, only some of them are the real boundaries of the desired object. Under these conditions, gray level based methods may easily converge to wrong boundaries, resulting in over or under segmentation. This could be corrected by refining the results through a manual work or via the implementation of other methods.

In contrast, structure based methods [12-15] are capable of dealing with unclear liver boundaries by utilising prior knowledge. Structure based methods can handle some problems which the gray level based methods cannot deal with. However, these methods require a great deal of training data to span all plausible conditions of the liver. Applying these methods incur significant difficulty when handling nonstandard shapes for the liver. In other words, it is very hard to develop a unified segmentation models for liver based on structure based methods. Instead of using gray level or shapes, texture based methods [16-19]utilises pattern recognition and machine learning to locate boundaries. As a result, these methods enable one to collectively consider more features. Texture based methods can produce better results when the boundaries are not clear. An accurate description of texture feature constitutes the main challenges in these methods in addition to the need for training data. Although there exist many descriptors, they are not like those described by human. Furthermore, selecting a proper descriptor out of many, poses another

problem. It should be noted that both machine learning and pattern recognition are still developing technologies with much weaker information processing abilities than human brain and they are not able to achieve good segmentation result without the use of other refined methods. In general, gray level based methods are more highly developed. They can conveniently be used to deal with complex segmentation tasks.

In the majority of cases, gray level based methods can attain satisfactory segmentation results. Structure based methods relies heavily on the shape of the object, a property making them a more robust technique. Finally, texture based methods attempt to emulate the procedure that human's brain processes. Level set methods have been investigated and widely utilized in image segmentation especially for medical images segmentation[20-22].Current approaches in using level set methods represent promising approaches for segmenting irregular object shapes such as liver. However it has a strict requirement on the initial position. It achieves a good result when the initial contour is placed near the target.3D liver segmentation methods can be categorized into two classes: direct 3D segmentation and propagation of the 2D slice-based segmentation. In terms of the first class, the user initializes a 3D deformable surface in multiple 2D slices of the liver, and then the initial 3D mesh is automatically refined by forces characterized by the image gradient and smoothness of the contour. This kind of method is time consuming and requires many user interactions that can lead to observer variability. The second class of 3D liver segmentation makes use of the slice-based propagation approach. In this technique, the 3D CT images are re-sliced into a number of 2D slices. A 2D segmentation is used in each slice, which is initialized by a propagated boundary from the previous 2D slice. This technique reduced a 3D segmentation problem to a sequence of 2D segmentation problems. Each of the reduced 2D segmentation sub-problems is much simpler than the original 3D segmentation problem, and it is also much cheaper on computational expense to incorporate 2D shape information as a shape constraint into the 3D segmentation procedure. Since the difference between adjacent slices is small, the final contour of one slice can provide useful information about the initial contour position and prior intensity and shape information which in turn enhances the segmentation performance of the level set method for the following slices. In this paper, we will modify the distance regularization level set [23] (DRLS) model in order to segment the liver contour in each 2D slice by using a new balloon forces that controls the direction of the evolution and slows down the evolution process in the region with weak or without edges. This papers organized as follows: Section 2 reviews the DRLS model. Section3 describes our methodology. Section4 shows some experimental results, and finally the conclusion is conveyed in Section 5.

2. DISTANCE REGULARIZED LEVEL SET METHOD

Li*et al.*[23]proposed a level set method termed as Distance Regularized Level Set (DRLS) model. The DRLS model uses an edge-based active contour method to drive the level set function (LSF) to the desired boundary, and provides a simple and efficient narrowband implementation without re-initialization.

Let $\phi : \Omega \to \Re$ be a level set function defined on domain Ω. An energy function $\varepsilon(\phi)$ is defined as:

$$\varepsilon(\phi) = \beta R_p(\phi) + \varepsilon_{ext}(\phi) \tag{1}$$

where $\beta > 0$ is a constant and $R_p(\phi)$ is the level set regularization term, defined by

$$R_p(\phi) = \int_{\Omega} p(|\nabla\phi|)dx \tag{2}$$

where p signifies an energy density function $p:[0,\infty) \to \Re$, defined as

$$p(s) = \begin{cases} \frac{1}{2}(s-1)^2, & \text{if } s > 1 \\ \frac{1}{(2\pi)^2}(1-\cos(2\pi s)), & \text{otherwise} \end{cases} \tag{3}$$

The minimization of the energy $\varepsilon(\phi)$ can be achieved by solving a level set evolution equation. For a LSF, an external energy function is defined by

$$\varepsilon_{ex}(\phi) = \lambda L_g(\phi) + \alpha A_g(\phi), \tag{4}$$

where λ and α are the coefficient of the length term $L_g(\phi)$ and area term $A_g(\phi)$, which is given by

$$L_g(\phi) = \int g\delta_\varepsilon(\phi)|\nabla\phi|dx \tag{5}$$

And

$$A_g(\phi) = \int gH(-\phi)dx \tag{6}$$

where $g \in [0,1)$ is an edge indicator function given by

$$g = \frac{1}{1+|\nabla G_\sigma \cdot I|^2}, \tag{7}$$

where G_σ is a Gaussian kernel with standard deviation σ, and I is the input image. The Dirac delta function δ_ε and Heaviside function H_ε in Eqs. (5)and (6) are approximated by the following smooth function δ_ε and H_ε, respectively, as in many level set methods:

$$\delta_\varepsilon(s) = \begin{cases} \frac{1}{2\varepsilon}[1+\cos(\frac{\pi t}{\varepsilon})], & \text{if } |t| \le \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

and

$$H_\varepsilon(s) = \begin{cases} \frac{1}{2}(1+\frac{t}{\varepsilon}+\frac{1}{\pi}\sin(\frac{\pi t}{\varepsilon})), & \text{if } |t| \le \varepsilon \\ 0, & \text{if } t < -\varepsilon \\ 1, & \text{if } t > \varepsilon \end{cases} \tag{9}$$

where ε is a constant, typically set to 1.5.The length term $L_g(\phi)$ was first introduced by Caselles*et al.*[24]in their proposed geodesic active contour (GAC) model. $L_g(\phi)$ computes the line integral of the function *g*along the zero level contour of ϕ, which is minimized when the zero level set of ϕ is located at the object boundaries which in turn keeps the curve smooth during the deformation.The area term $A_g(\phi)$ calculates the weighted area inside the evolving contour. It is introduced to speed up the motion of the zero level contourwhen the contour is far away from the desired object boundaries and slow down the expanding and shrinking of the zero level contour when it arrives at object boundaries where *g* is smaller. $A_g(\phi)$ represents a balloon forces in which the sign of α controls the direction of the level set evolution (shrinking or expanding).The level set evolution equation in the DRLS model is defined by:

$$\frac{\partial \phi}{\partial t} = \beta.\mathrm{div}(d_p(|\nabla \phi|)\nabla \phi) + \delta(\phi).\lambda.\mathrm{div}\left(g\frac{\nabla \phi}{|\nabla \phi|}\right) + \delta(\phi).\alpha.g \tag{10}$$

The problem with the DRLS model in the case of liver segmentation is that the curve will evolve and deviate from the liver boundary in the region with weak or without edges. In this contribution, we will modify the distance regularization level set method [23] (DRLSM) by adding a new balloon force to guide the evolution process and discourage the evolving contour from leaking at a region with a weak or without an edges and from going far from the liver boundary.

3. THE PROPOSED MODEL

In this paper, we propose a new balloon force that controls the direction of the evolution and slows down the evolving contour at weak or blurred edges. Since the liver has a very similar intensity with its adjacent organs, this could easily result in over and/or under segmentation results. The DRLS model does not perform well with liver segmentation. We will modify the DRLS model to segment the liver contour in each 2D slice by using a new balloon force that controls the direction of the evolution and slows down the evolution process in the region with weak or without edges, which subsequently discourage the evolving contour from leaking at a region with a weak or without an edge and from deviating from the liver boundary. Our balloon term will be built using the probability density function. The methodology encompasses steps described in the following sections.

3.1 PRE-PROCESSING

The intensity distribution of the liver is irregular due to noises, so liver segmentation without pre-processing is difficult. A smoothing step, in theory, would make the intensity distribution less variable. In our work, a Gaussian filter is used as a smoothing step.

3.2 SEGMENTATION OF THE REFERENCE SLICE

This step is the most important step in our 3D liver segmentation method. The segmented liver contour will be the initial contour for the adjacent slice so the segmentation result should be

accurate. The starting slice or the reference slice can be selected as a middle or the largest slice of the liver volume. In this contribution we used the Active Shape Model (ASM) [25] to segment the reference slice.

3.3 2D SLICE BASED PROPAGATION APPROACH

Since the variation of shape and intensity between the adjacent slices are very small we can use these information from the previous slice to segment the next slice. In our method we compute the mean intensity μ and the variance σ of the segmented slice. According to [26],about 98% of liver pixel is located in $[\mu\text{-}3\sigma, \mu+3\sigma]$.Generating an evolution region by expanding the previous segmented slice by a number of pixels and computing the probability density function inside this region using the following equations:

$$B(X) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, & \text{if } x \in [\mu - 3\sigma, \mu + 3\sigma] \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

We then shrink the previous segmented slice and use it as the initial contour for its adjacent slices in both directions.

3.4 MODIFIED DRLS MODEL

Since the liver boundary to be segmented is not far from the contour propagated from the previous slice, a shape and intensity prior information will discourage the evolving contour from leaking at a region with a weak edge or without an edge. We have modified the DRLSM by adding the probability density energy term to the evolution equation and use it as a balloon forces to control the direction and the speed of the evolution process

$$E(\phi) = \rho.R_P(\phi) + \lambda.L_g(\phi) + \alpha.B(\phi) \tag{12}$$

where ρ, λ and α are the coefficients of regularization term, length term, and probability term, respectively. This energy functional can be minimized by solving the following gradient flow:

$$\frac{\partial\phi}{\partial t} = \rho.\mathrm{div}(d_p(|\nabla\phi|)\nabla\phi) + \delta_\varepsilon(\phi).\lambda.\mathrm{div}\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \delta_\varepsilon(\phi).\alpha.B(\phi) \tag{13}$$

The above procedure is repeated until the contours in all 2D slices of the 3D image are segmented. A 3D liver surface is reconstructed from the contours segmented from all 2D slices.

4. RESULT AND DISCUSSION

In the DRLS model, two segmentation stages are applied. The first stage is for evolving the contour in the direction of the object boundary, speeding up the evolution process when the evolving contour is far from the object boundary and slowing down the evolution process when the evolving contour is close to the object boundary. The second stage concerns with the

refinement of the segmentation results. In each experiment, we selected values of ρ , λ and α to be 0.02, 5 and -1 for the first stage and 0.02, 5 and 0 for the second stage, respectively. The zero level set is initialized as a binary function and evolves according to the evolution equation, Eq. (13) for our model and Eq. (10) for the DRLS model.

Figures 1 and 2 present segmentation results of the DRLS model and the proposed model in a liver CT slice. Our model performs well and gives a satisfactory result comparing to the DRLS model. The DRLS model fails to segment the liver boundary and the evolving contour leaks from the region with weak edges. Our balloon force slows down the evolution process close to the liver boundary and stops the evolving contour from going far in the region with weak or without edges. Comparing with the DRLS model, our model is more effective in dealing with over segmentation problem.

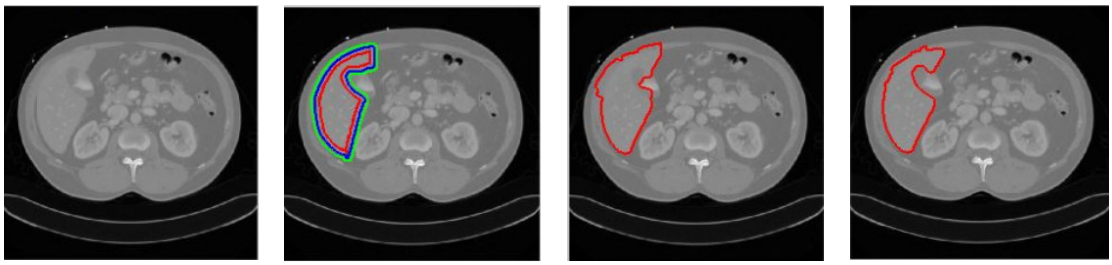


Figure1. Comparisons of liver segmentation result between the DRLS model and the proposed method. The first image shows the liver slice in a CT scan; the second image shows the previous segment contour in blue, evolving region in green and the initial contour in red. The third image shows the final segmentation result of the DRLS model and the fourth image shows the final segmentation result with our proposed method.

Figure2. Comparisons of liver segmentation result between the DRLS model and the proposed method. The first image shows the liver slice in a CT scan; the second image shows the ground truth segmented manually by a radiologist. The third image shows the final segmentation result of the DRLS model and the fourth image shows the final segmentation result with our proposed method.

Figure 3 shows some examples of liver extraction results based on our proposed method. We tested our model on a liver dataset containing 10 volumes of abdominal CT images. Each volume has 64 slices and the size of each slice is 512x512pixels. Each slice in the dataset is provided with corresponding ground truth segmented manually by a radiologist. The model deals very well with over segmentation problem. Our model can handle the over segmentation problems very well in comparison with the DRLS model that is not able to carry out this task.

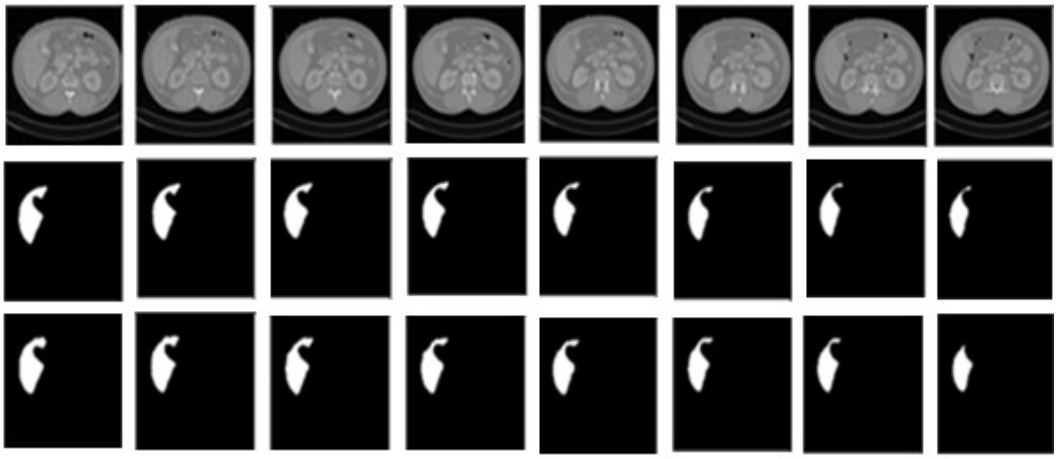


Figure 3.The experimental result of our proposed method on a sequence of liver slices for one person. The first row shows the liver slice in a CT scan; the second row shows the ground truth segmented manually by the radiologist. The third row shows the final segmentation result of our proposed method.

5. CONCLUSION

A novel balloon force method is presented herein. The main merits of this approach lies in its ability to guide the direction of the evolving contour and slows down the evolving contour in regions that are associated with weak or without edges and discourages the evolving contour from going far away from the liver boundary or from leaking at a region that has a weak edge, or does not have an edge. The model utilises a modified Distance Regularized Level Set (DRLS) model. The experimental results demonstrate that the method can achieve a satisfactory result. Our model proves to be more effective in dealing with over segmentation problems if compared with the DRLS model.

REFERENCE

[1] Luo, S., Li, X. & Li, J. (2014) "Review on the Methods of Automatic Liver Segmentation from Abdominal Images",Journal of Computer and Communications. Vol. 2, No.2, pp 1-7.

[2] Adams, R. & Bischof, L. (1994) "Seeded region growing",IEEE Transactions onPattern Analysis and Machine Intelligence, Vol. 16, No.6, pp 641-647.

[3] Beck, A. & Aurich, V. (2007) "Hepatux–a semiautomatic liver segmentation system",3D Segmentation in The Clinic: A Grand Challenge. pp 225-233.

[4] Pohle, R.,& Toennies, K. D. (2001)Segmentation of medical images using adaptive region growing. in Medical Imaging International Society for Optics and Photonics.

[5] Mortelé, K. J., Cantisani, V., Troisi, R., De Hemptinne, B.,& Silverman, S. G. (2003) "Preoperative liver donor evaluation: imaging and pitfalls",Liver Transplantation. Vol. 9, No.9, pp S6-S14.

[6] Kumar, S., Moni, R.,& Rajeesh, J. (2013) "Automatic liver and lesion segmentation: a primary step in diagnosis of liver diseases",Signal, Image and Video Processing. Vol. 7, No.1, pp 163-172.

[7] Platero, C., Poncela, J. M., Gonzalez, P., Tobar, M. C., Sanguino, J., Asensio, G. & Santos, E. (2008)Liver segmentation for hepatic lesions detection and characterisation. in 5th IEEE International Symposium onBiomedical Imaging: From Nano to Macro, IEEE.

[8] Oliveira, D.B., Feitosa, R. Q. & Correia, M. M. (2009)Liver Segmentation using Level Sets and Genetic Algorithms. in VISAPP (2).

[9] Yang, H., Wang, Y., Yang, J.,& Liu, Y. (2010)A novel graph cuts based liver segmentation method. in International Conference onMedical Image Analysis and Clinical Applications (MIACA), IEEE.

[10] Chen, Y.-W., Tsubokawa, K.,. & Foruzan, A. H., (2010)Liver segmentation from low contrast open MR scans using k-means clustering and graph-cuts, in Advances in Neural Networks-ISNN 2010, Springer. pp 162-169.

[11] Foruzan, A. H., Yen-Wei, C., Zoroofi, R. A., Furukawa, A., Masatoshi, H. & Tomiyama, N. (2013) "Segmentation of Liver in Low-Contrast Images Using K-Means Clustering and Geodesic Active Contour Algorithms",IEICE Transaction on Information and Systems, Vol. 96, No.4, pp 798-807.

[12] Liu, J. & Udupa, J. K. (2009) "Oriented active shape models",IEEE Transactions onMedical Imaging, Vol. 28, No.4, pp 571-584.

[13] Heimann, T., Wolf, I. & Meinzer, H.P. (2006)Active shape models for a fully automated 3D segmentation of the liver–an evaluation on clinical data, in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, Springer. pp 41-48.

[14] Erdt, M., Steger, S., Kirschner, M. & Wesarg, S. (2010)Fast automatic liver segmentation combining learned shape priors with observed shape deviation. in IEEE 23rd International Symposium onComputer-Based Medical Systems (CBMS), IEEE.

[15] Badakhshannoory, H. & Saeedi, P. (2011) "A model-based validation scheme for organ segmentation in CT scan volumes",IEEE Transactions onBiomedical Engineering, Vol. 58, No.9, pp 2681-2693.

[16] Huang, W., Tan, Z., Lin, Z., Huang, G., Zhou, J., Chui, C., Su, Y. & Chang, S. (2012)A semi-automatic approach to the segmentation of liver parenchyma from 3D CT images with Extreme Learning Machine. in Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE.

[17] Danciu, M., Gordan, M., Florea, C. & Vlaicu, A. (2012)3D DCT supervised segmentation applied on liver volumes. in 35th International Conference onTelecommunications and Signal Processing (TSP), IEEE.

[18] Luo, S., Li, X. & Li, J. (2013) "Improvement of Liver Segmentation by Combining High Order Statistical Texture Features with Anatomical Structural Features",Engineering. Vol. 5, No.5, pp 67-72.

[19]Luo, S., Hu, Q., He, X. , Li, J. , Jin, J. S. & Park, M. (2009)Automatic liver parenchyma segmentation from abdominal CT images using support vector machines. in ICME International Conference onComplex Medical Engineering, IEEE.

[20] Xu, C., Pham, D. & Prince, J. (2000)Medical Image Segmentation Using Deformable Models, in SPIE Handbook on Medical Imaging J.M. Fitzpatrick and M. Sonka, Editors, pp 129-174.

[21] Altarawneh, N. M., Luo, S., Regan, B., Sun, C. & Jia, F. (2014) "Global threshold and region-based active contour model for accurate image segmentation",Signal & Image Processing: An International Journal (SIPIJ). Vol.5, No.3, pp 1-11.

[22] Altarawneh, N. M., Luo, S., Regan, B., Sun, C. (2014) "A novel global threshold-based active contour model",Second International Conference on Signal, Image Processing and Pattern Recognition(SIPP),pp 245-254.

[23] Li, C., Xu, C., Gui, C. & Fox, M. D. (2010) "Distance regularized level set evolution and its application to image segmentation",IEEE Transactions onImage Processing, Vol. 19, No.12, pp 3243-3254.

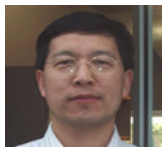
[24] Caselles, V., Kimmel, R. & Sapiro, G. (1997) "Geodesic active contours",International Journal of Computer Vision. Vol. 22, No.1, pp 61-79.

[25] Van Ginneken, B., Frangi, A. F., Staal, J. J., Ter Haar Romeny, B. M. & Viergever, M. A. (2002) "Active shape model segmentation with optimal features",IEEE Transactions onMedical Imaging, Vol. 21, No.8, pp 924-933.

[26] Li, X., Luo, S. & Li, J. (2013) "Liver Segmentation from CT Image Using Fuzzy Clustering and Level Set",Journal of Signal and Information Processing. Vol. 4, No.3, pp 36-42.

AUTHORS

Nuseiba Altarawneh is a PhD student at the University of Newcastle, Australia. Her research interests include computer vision, image analysis, and pattern recognition. She obtained the M.E. degree in Computer Science from the University of Jordan in 2009.



Dr. Suhuai Luo received the PhD degree in Electrical Engineering from the University of Sydney, Australia in 1995. From 1995 to 2004, he worked as a senior research scientist with the Commonwealth Scientific and Industrial Research Organization Australia and the Bioinformatics Institute Singapore. He is currently a senior lecturer with the University of Newcastle, Australia. His research interest is in information technology and multimedia, including health informatics, machine learning, image processing, computer vision, and Internet-oriented IT applications.

Dr. Brian Regan is a senior lecturer in IT at the University of Newcastle. He is part of the Applied Informatics Research Group (AIR) with interests in health informatics, visualization and development methodologies.

Dr. Changming Sun received the PhD degree in the area of computer vision from Imperial College London in 1992. Then, he joined CSIRO Computational Informatics, Australia, where he is currently a principal research scientist carrying out research and working on applied projects. His research interests include computer vision, image analysis, and pattern recognition. He has served on the program/organizing committees of various international conferences. He is an Associate Editor for EURASIP Journal on Image and Video Processing, a Springer One journal.

IMPROVED ALGORITHM FOR ROAD REGION SEGMENTATION BASED ON SEQUENTIAL MONTE-CARLO ESTIMATION

Zdenek Prochazka

Department of Information Engineering,
National Institute of Technology, Oita College, Oita, Japan
zdenek_p@oita-ct.ac.jp

ABSTRACT

In recent years, many researchers and car makers put a lot of intensive effort into development of autonomous driving systems. Since visual information is the main modality used by human driver, a camera mounted on moving platform is very important kind of sensor, and various computer vision algorithms to handle vehicle surrounding situation are under intensive research. Our final goal is to develop a vision based lane detection system with ability to handle various types of road shapes, working on both structured and unstructured roads, ideally under presence of shadows. This paper presents a modified road region segmentation algorithm based on sequential Monte-Carlo estimation. Detailed description of the algorithm is given, and evaluation results show that the proposed algorithm outperforms the segmentation algorithm developed as a part of our previous work, as well as an conventional algorithm based on colour histogram.

KEYWORDS

Lane Recognition, Road Region Segmentation, Sequential Monte-Carlo Estimation

1. INTRODUCTION

In recent years, many car makers put a lot of intensive effort into development of autonomous driving systems, and published their plans to start production of cars with self-driving ability within next few years.

Autonomous driving systems exploit information from various sensors, and interpret sensory information to identify appropriate navigation paths, as well as obstacles and relevant signage. Among various sensors, a camera is an indispensable one, because real human driver rely mainly on visual information, and transport infrastructure is most suited to the human vision system.

The development of computer vision techniques for autonomous driving is a very challenging task, and intensive research on this topic is carried out. Our work deals with an issue of computer vision based lane detection.

Several lane detection methods based on detection of road boundaries were already developed in the past [1–5]. Detection of road boundaries is well suited for roads with clear surface marking,

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 171–183, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41213

however it can easily fail if road boundaries are not clear, or if strong edges caused by non-road objects are present in image. It is also not obvious, how to handle complicated road shapes, splitting and merging of traffic line, complicated road marking, or sudden changes in road width.

To overcome the limitations of the boundary based approach, techniques for segmentation of road as a region are intensively studied. A common issue with the road region segmentation is how to build a classifier to make decision whether a image pixel belongs to the road surface or not. There are two basic approaches to this issue. The first one is adopting an algorithm from field of machine learning. Methods based on neural networks [6] or SVM [7,8] are reported. Since the training of the classifier is generally an off-line process, the main difficulty with such kind of methods is how to bring adaptability to new data which appear during detection stage and were not part of the original training set.

The second approach are methods based on idea of statistical decision. Statistical distribution of road features is modelled as parametric or non-parametric probability distribution function (pdf), and the pdf is used to decide whether a given image pixel is part of road or not. Methods using simple Gaussian colour model [9], Gaussian mixture model [10], histogram of colour components [11], histogram of illumination invariant features [12] have been reported. The main difficulty with the methods based on statistical decision is the fact, that it is not obvious how to estimate pdf. To estimate pdf for each incoming frame, we need to some guess where the road is, which can easily pose a chicken and egg problem. In many cases, some simple heuristics or ad-hoc solutions are adopted. Furthermore, although the road segmentation should be performed over an sequential image, many of the developed methods study segmentation only from single frame of road image, and behaviour for image sequence is not studied enough.

In our work, we have focused to approach of statistical decision, and attempt to solve estimation of pdf for sequential image in more systematic way. In our previous report, we have already proposed a road segmentation algorithm based on sequential Monte-Carlo (SMC) estimation [13], and applied it to region-based pathway estimation [14]. Although this segmentation method yields better results than the conventional one, and it is quite robust to parameter setting, it still doesn't produce optimal result under certain circumstances. This paper shows a limitation of the original algorithm and presents a new algorithm with improved segmentation ability. The proposed algorithm is based upon a sound theoretical framework, and evaluation results for various kinds of image data are shown to demonstrate effectiveness of our proposal.

Although an issue of shadows is not discussed in this paper, it should be mentioned that the proposed method can be easily adapted to other type of features, which are claimed to be less sensitive to illumination [10,12].

This document describes, and is written to conform to, author guidelines for the journals of AIRCC series. It is prepared in Microsoft Word as a .doc document. Although other means of preparation are acceptable, final, camera-ready versions must conform to this layout. Microsoft Word terminology is used where appropriate in this document. Although formatting instructions may often appear daunting, the simplest approach is to use this template and insert headings and text into it as appropriate.

2. SMC ESTIMATION EASED ROAD SEGMENTATION ALGORITHM

This section at first recapitulates the original SMC based road region segmentation algorithm, and discuss its limitations. After that a new algorithm with improved ability is proposed and described in details.

2.1. Former Algorithm

Let $\boldsymbol{f}$ be a vector of road features. In this work, a five dimensional feature vector $\boldsymbol{f} = (f_1, \cdots, f_5)^T$ is used, where f_1 and f_2 corresponds to pixel coordinates x and y respectively, and f_3, f_4, f_5 corresponds to colour components r, g, b. Although r, g, b were adopted, the algorithm is generally not limited to these features. For convenience we define also vectors $\boldsymbol{x} = (x, y)^T = (f_1, f_2)^T$ and $\mathbf{c} = (r, g, b)^T = (f_3, f_4, f_5)^T$ as shorthand for vector of coordinates and colour components respectively. With this notation, the feature vector can be expressed also as $\boldsymbol{f} = (\boldsymbol{x}^T, \boldsymbol{c}^T)^T$.

1: Initialise sample set $S^{(0)}$
2: **while** Image frame successfully captured **do**
3: Proceed discrete time k.
4: Draw randomly N samples from the sample set $S^{(k-1)}$, according to the importance weights $w_i^{(k-1)}$.
5: Apply transition model to each drawn sample.

$$\boldsymbol{f}_i^{(k)} = \boldsymbol{f}_i^{(k-1)} + \boldsymbol{\nu}_i^k \tag{2}$$

6: Set new weight for each sample, using an observation pdf.

$$w_i^{(k)} = p(\boldsymbol{z}_i^{(k)} | \boldsymbol{c}_i^{(k)}) \tag{3}$$

7: Normalise weights to be

$$1 = \sum_{i=1}^{N} w_i^{(k)} \tag{4}$$

8: Replace $S^{(k-1)}$ by $S^{(k)}$.
9: Use the estimation result to classify road pixels.
10: **end while**

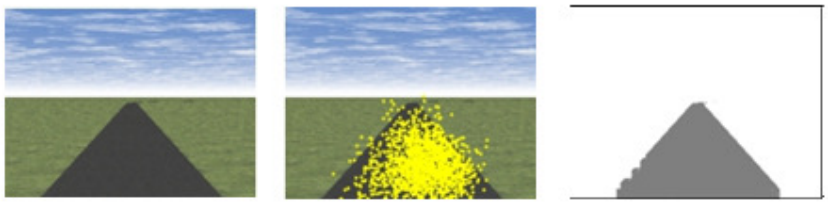
Figure 1. Original Algorithm

The goal is to estimate pdf of road features $p(\boldsymbol{f})$, and use the estimated pdf to classify road and non-road pixels. Since the road can take various shapes, $p(\boldsymbol{f})$ is expected to be a non-Gaussian distribution at least with respect to the $\boldsymbol{x}$. This suggests that we need an estimation method which can handle general type of distribution. Since the road usually looks very similar between two successive frames, and overall visual appearance changes relatively slowly compared to frame rate, temporal continuity can be involved into the pdf estimation, and it should be possible to model a visual changes of road as a stochastic process.

Taking the above mentioned into account, we have focused to SMC (Sequential Monte-Carlo) estimation as a basic tool for estimation of the p (f). The original segmentation algorithm was inspired by CONDENSATION algorithm [15]. The pdf is expressed in discrete form as a set of weighted samples S ,

$$S = \{(\boldsymbol{f}_1, w_1), (\boldsymbol{f}_2, w_2), \Lambda\ (\boldsymbol{f}_N, w_N)\} \tag{1}$$

where $\boldsymbol{f}_i$ is a sample of the feature vector $\boldsymbol{f}$ and w_i is an importance weight assigned to the sample. Estimation is performed by repetitive update of the S according to the algorithm shown in Figure 1. The superscript k stands for discrete time here. The $\boldsymbol{v}$ in the eq.(2) means a vector of Gaussian random numbers with zero mean and variances $\sigma_1,\Lambda\ \sigma_5$. The $p(\boldsymbol{z}_i \mid \boldsymbol{c}_i)$ in eq.(3) is an observation pdf defined by eq.(5).



(a) Original Image (b) Samples (c) Segmentation Result

Figure 2: Example of Segmentation Result

$$p(\boldsymbol{z}_i \mid \boldsymbol{c}_i) = \exp\left(-\frac{1}{2}(\boldsymbol{c}_i - \boldsymbol{z}_i)^T \Sigma_m^{-1}(\boldsymbol{c}_i - \boldsymbol{z}_i)\right) \tag{5}$$

The symbol Σ_m here means a diagonal matrix of variances, controlling width of exponential function for each colour component, and $\boldsymbol{z}_i$ means a colour component observed at position obtained from eq.(2). Simply said, the $p(\boldsymbol{z}_i \mid \boldsymbol{c}_i)$ is a measure of similarity between colour components $\boldsymbol{c}_i$ predicted by eq.(2), and colour components observed at predicted positions. The calculated similarity is than assigned as an importance weight.

The above described algorithm estimates the pdf in a form of weighted samples. To evaluate pdf for an arbitrary pixel, we perform a smoothing of the estimated pdf, in a way similar to non-parametric density estimation, using kernel function ϕ. Hence, the pdf $p(\boldsymbol{f})$ for an arbitrary pixel values $\boldsymbol{f}$ is calculated as

$$p(\boldsymbol{f}) = \sum_{i=1}^{N} w_i \phi(\boldsymbol{f}, \boldsymbol{f}_i) \tag{6}$$

$$\phi(\boldsymbol{f}, \boldsymbol{f}_i) = \beta \exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{f}_i)^T \Sigma_s^{-1}(\boldsymbol{f} - \boldsymbol{f}_i)\right)$$

where the symbol Σ_s is a diagonal matrix of parameters, controlling window width for each component of $\boldsymbol{f}$, and the β is a normalising constant. Segmentation of the road region is performed by evaluation of $p(\boldsymbol{f})$ for each pixel, followed by thresholding of obtained values to decide road and non-road pixels.

We have already reported that the above described segmentation method is quite robust to parameter tuning and yield better results than conventional segmentation methods like region growing or simple Gaussian colour model [16]. However there are still issues to be solved. Let's look at Figure 2 to for explanation.

The image shown in Figure 2(a) is a synthetic image, and the road region can be segmented very easily here.The Figure 2(b) shows a spatial distribution of samples $\boldsymbol{f}_i$, after the method described above has been applied. As can be seen, the samples are not distributed uniformly, and only few samples appear over left and right side of low part of the road region. As a consequence, these parts are not segmented properly, as is shown in Figure 2(c). Such improper distribution of samples can cause serious problems in proper segmentation of roads with multiple traffic lines.

There is another issue, which can't be simply shown by static image. Since the estimation of pdf of the road features is treated as an stochastic process, certain noise has to be introduced into the model. Due to this stochastic nature, the positions of samples change between subsequent frames, and segmented parts of road covered only by few of samples are very sensitive to such changes. Due to this, rapid changes in shape of the segmented road appear from frame to frame, although the road doesn't changes significantly in the image. We refer these quick temporal changes of segmentation result as fluctuations in the following. It is clearly an unwelcome phenomenon which have to be eliminated.

After careful revision of the original algorithm, we have proposed a new algorithm, which attempts to solve the above mentioned issues.

2.2. Proposed Algorithm

The limitations of the original algorithm are caused by improper distribution of the samples during the estimation process. To obtain proper segmentation result, samples should be spread uniformly over the road region, and the sampling method described in step 4 of the Figure 1 can't guarantee this property. In the original algorithm, the only way to spread samples more widely over the road region is to increase amount of transition noise added to spatial components. However this cause larger fluctuations of the segmented region, therefore limitations of the original algorithm cannot be simply solved by parameter tuning.

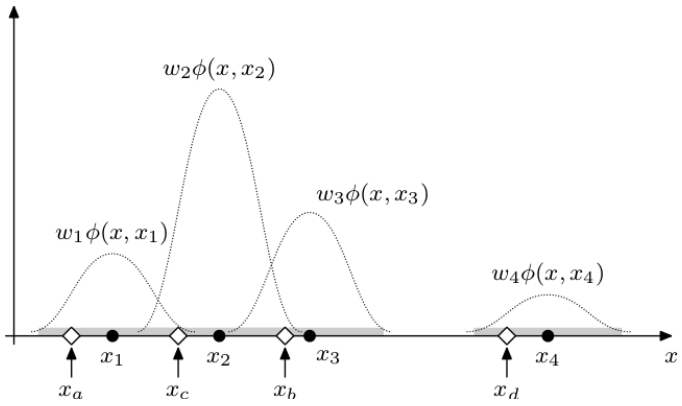


Figure 3: Assignment Between Subsequent Samples

Our idea to solve discussed issues is based on the principle of importance sampling [17]. In this framework, it is supposed that samples can be easily drawn from some probability distribution function $q(\boldsymbol{f}_i^{(k)} \mid \boldsymbol{f}_i^{(-1)}, \boldsymbol{z}_i^{(k)})$, which is called importance density. For samples generated from the importance density, an update of the sample weights is done according to the following form.

$$w_i^{(k)} \propto w_i^{(k-1)} \frac{p(z^{(k)} \mid f_i^{(k-1)}) p(f_i^{(k)} \mid f_i^{(k-1)})}{q(f_i^{(k)} \mid f_i^{(k-1)}, z^{(k)})} \tag{7}$$

Although principle of importance density provides sound theoretical framework, the specific forms of $p(z^{(k)} \mid f_i^{(k-1)})$, $p(f_i^{(k)} \mid f_i^{(k-1)})$ and $q(f_i^{(k)} \mid f_i^{(-1)}, z_i^{(k)})$ are not obvious, and strongly depends on particular application.

In our case, we have chosen $q(f_i^{(k)} \mid f_i^{(-1)}, z_i^{(k)})$ as a uniform distribution with respect to spatial components x , where the range of x is limited to region segmented at the previous time step. Simply said, spatial components of the generated samples x_i $(i = 1, \Lambda, N)$ are uniformly drawn from the region obtained at $k-1$.

To evaluate $p(f_i^{(k)} \mid f_i^{(k-1)})$, we need to establish relations between old samples and new samples generated according to importance density. This is not straightforward, and our idea is to perform assignment in a way described in Figure 3 for single variate case. Let the x_1, Λ, x_4 be old samples, and let the intervals highlighted by light grey be segmented parts of x axis. We draw new samples from the light grey intervals randomly, according to a uniform distribution. New samples are shown by diamond and the alphabet subscript here. For each new sample x_i $(i = a, b, c, d)$ we assign one of the old samples x_j $(j = 1,2,3,4)$, using the weighted kernel function ϕ, already mentioned in eq.(6), as an assignment criteria.

$$l = \arg \max_{j=1,\Lambda,4} (w_j \phi_x(x_i, x_j)) \tag{8}$$

Here, ϕ_x stands for univariate Gaussian kernel function. The l obtained by eq.(8) is than used as a subscript of assigned old sample. For the situation in Figure 3, the resulting assignment is (x_a, x_1), (x_b, x_3), (x_c, x_2), (x_d, x_4).

We have described the basic idea of sample assignment for simple univariate case. Next we explain, how to extend this idea to the our case of f. Since our importance density is defined in spatial subdomain x, we perform assignment of samples with respect to x. Therefore, after the new samples $x_i^{(k)}$ are generated according to importance density, we assign to each generated sample an old one, for which the $w_j^{(k-1)} \phi_x \left(x_i^{(k)}, x_j^{(k-1)} \right)$ is maximal. The symbol ϕ_x here stands for kernel function of the same form as in eq.(6), however applied only to spatial components

.

By the above described procedure, we can generate spatial part $x_i^{(k)}$ of the new samples, and perform assignment between new and old samples. However, we said nothing about the colour part $c_i^{(k)}$ so far. To generate colour part of the new samples, and build up complete samples $f_i^{(k)}$, we simply apply transition equation to colour part of assigned old sample.

$$c_i^{(k)} = c_l^{(k-1)} + {}_c v_i \tag{9}$$

Here the ${}_c v_i$ means a colour part of v from the eq.(2), and the subscript l means a label of an assigned old sample.

After the new sample set $\boldsymbol{f}_i^{(k)}$ is completed, we evaluate transition pdf in weight update equation as

$$p(\boldsymbol{f}_i^{(k)} \mid \boldsymbol{f}_l^{(k-1)}) = \phi(\boldsymbol{f}_i^{(k)} \mid \boldsymbol{f}_l^{(k-1)}) \tag{10}$$

where ϕ has a same form as in eq.(6), and l is a label of sample assigned to the i-th one.

The last component needed for weight update calculation in eq.(7) is the observation pdf $p(\boldsymbol{z}^{(k)} \mid \boldsymbol{f}_i^{(k-1)})$, and it has the same form as was already shown by eq.(5).

1: Initialise sample set $S^{(0)}$.
2: Segment initial region.
3: **while** Image frame successfully captured **do**
4: Proceed discrete time k.
5: Draw randomly N samples of $\boldsymbol{x}_i^{(k)}$, uniformly from segmented road region.
6: Assign $\boldsymbol{x}_l^{(k-1)}$ to each $\boldsymbol{x}_i^{(k)}$
7: Apply transition model to colour part of each assigned old sample.

$$\boldsymbol{c}_i^{(k)} = \boldsymbol{c}_l^{(k-1)} + {}_c\boldsymbol{\nu}_i \tag{11}$$

8: Set new weight for each sample, using transition and observation pdfs.

$$w_i^{(k)} = w_l^{(k-1)} p(\boldsymbol{z}^{(k)} | \boldsymbol{f}_i^{(k)}) p(\boldsymbol{f}_i^{(k)} | \boldsymbol{f}_l^{(k-1)}) \tag{12}$$

9: Normalise weights to be

$$1 = \sum_{i=1}^{N} w_i^{(k)} \tag{13}$$

10: Replace $S^{(k-1)}$ by $S^{(k)}$.
11: Use the estimation result to classify road pixels.
12: **end while**

Figure 4. Proposed Algorithm

The proposed algorithm is summarised in Figure 4.In the weight update eq.(12) of the step 8, the value of importance density q is omitted, since we draw samples according to uniform distribution.

At the end of this section, we show a simple example of the segmentation result obtained by the proposed algorithm. The result for the same data as in Figure 2 is shown in Figure 5. As can be seen, samples are spread over all road region and there are no missing parts of segmented road region.

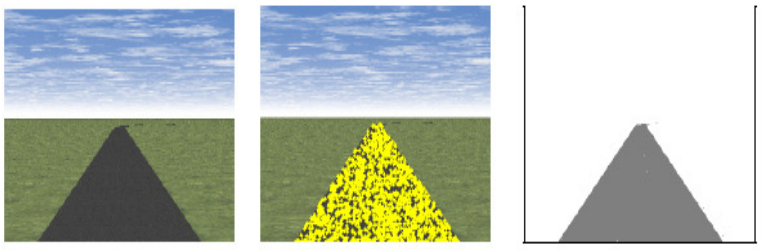


Figure 5: Example of Segmentation Result

In the next section, we present more detailed evaluation of the proposed algorithm for various kinds of image data.

3. EVALUATION OF THE PROPOSED ALGORITHM

The evaluation of the proposed algorithm have been performed on the following kinds of image data.

A: Sequences of Still Images

Approximately 200 images, capturing frontal roads viewed from car, were randomly collected. Half of them are structured roads, second half are unstructured and rural roads. For the each image, a sequential image was created by repeating of the same still image. Created sequential images shows snapshots of real road, but since there are no temporal changes between subsequent frames, these data can be used to assess fluctuations of the resulting segmentation.

B: Synthetic Image Sequence

Sequence of rendered CG images, already used in Figure 2 and Figure 5, was created. The sequence simulates a vehicle moving by 60 km/h on road with 4 curves (R=150m, 100m, 80m, 50m). Vehicle steering is simulated by changes in lateral position and pitch angle. Since the sequence is composed of road images with almost homogeneous colour, the perfect segmentation of road region can be done very easily, and this sequence can be used to assess quality of segmentation.

C: Real Image Sequences

Real image sequences, already mentioned in our report [14], with representative frames shown in Figure 6. These sequences are real data, therefore they can be used to assess overall ability of the algorithm.

For all above mentioned data, ideally segmented road regions were prepared as ground truth. The segmentation was done manually for data A and C, and automatically during CG rendering for data B.

As a quantitative measure of the segmentation ability, we adopted Jaccard index J, defined by eq.(14).

$$J = \frac{|R_s \cap R_g|}{|R_s \cup R_g|} \tag{14}$$

Here, R_s is a set of road pixels obtained by segmentation algorithm, and R_g is a set of road pixels for ground truth. Symbol $|\ |$ means a set size here. Jaccard index J takes value 1 if both sets R_s and R_g are identical.

Besides the SMC based algorithms described in this paper, we implemented another segmentation algorithm based on seeded region growing (SRG). Overall configuration is shown in Figure 7.

Figure 6: Real Image Sequences

SRG based algorithm use normalised histogram of features as a merging criteria (pixel is merged if histogram value for pixel features is greater than certain threshold). The histogram of features is calculated from region placed over low part of road image. This follows the same strategy as I reports [11] and [12]. Placement of seed points follows suggestions given by report [12].

SRG based algorithm, original SMC based algorithm and proposed SMC based algorithm were evaluated for data A, B, C. Before evaluation, all necessary parameters were once tuned to maximise average value of Jaccard index eq.(14). After tuning, parameter values were fixed, and same fixed values were used for all tested data.

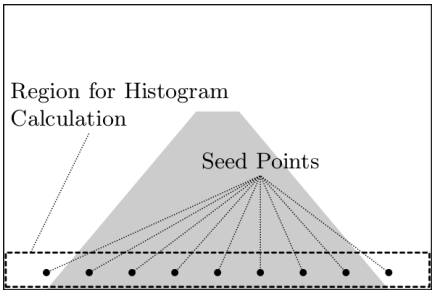
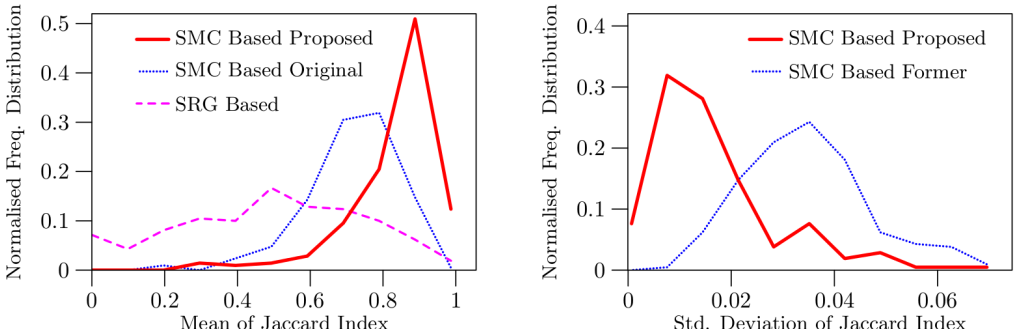


Figure 7: Configuration of SRG Based Reference Algorithm

3.1. Evaluation for Sequences of Still Images

Three mentioned algorithms, were applied to data A, and Jaccard index eq.(14) was calculated for segmentation results of each particular sequence frame by frame. Next, average μ_J and standard deviation σ_J of Jaccard index was calculated separately for each sequence, and these values were used to form histograms of μ_J and σ_J respectively. The resulting histograms are shown in Figure 8. Since data A have no temporal changes, results by SRG based algorithm contains no fluctuations, and evaluation of σ_J has no meaning for data A. Therefore, only result for SMC based algorithm is shown in Figure 8(b). Ideal values of μ_J and σ_J are $\mu_J = 1$, $\sigma_J = 0$, where higher μ_J indicates better segmentation result, and lower σ_J indicates less amount of fluctuations. Figure 8(a) shows that peak of histogram for the proposed algorithm is closer to

ideal value $\mu_J = 1$, and (b) of the same figure shows that peak of histogram for the proposed algorithm is closer to ideal value $\sigma_J = 0$. It means, that for majority of the data A, the proposed algorithm yields segmentation results closer to ground truth, while fluctuations of segmented regions are reduced. The result for SRG based algorithm is surprisingly wrong. The values of average of Jaccard index are distributed in wide range, which shows that the SRG based algorithm is not robust to parameter setting.

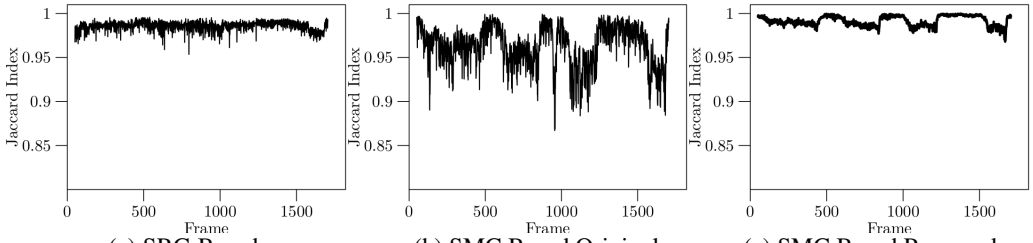


(a) Histogram of Means μ_J (b) Histogram of Standard Deviations σ_J

Figure 8: Histogram of Jaccard Index for Sequences Still Image

3.2. Evaluation for Synthetic Image Sequence

Three mentioned algorithms were applied to data B, and Jaccard index eq.(14) was calculated for each segmented frame of image sequence. Graphs of the Jaccard index are shown in Figure 9. It can be seen, that proposed algorithm yields similar result as SRG based algorithm, where result in Figure 9(a) is more noisy than Figure 9(c). This is caused by fact, that SRG based algorithm segments each frame as an independent still frame and there is no mechanism for temporal continuity. The result for original algorithm in Figure 9(b) contains large fluctuations, and magnitude of fluctuations is significantly reduced by the proposed algorithm.



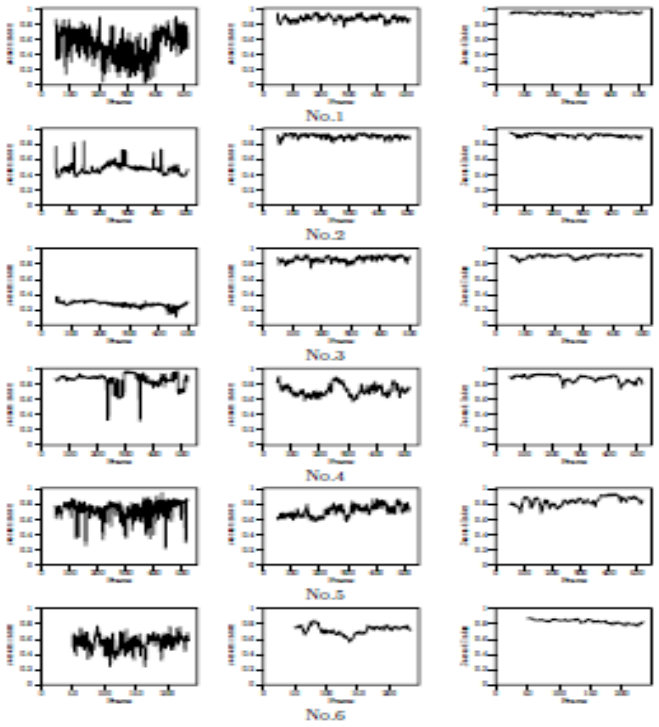
(a) SRG Based (b) SMC Based Original (c) SMC Based Proposed

Figure 9: Jaccard Index for Synthetic Image Sequence

3.3. Evaluation for Real Image Sequences

Three mentioned algorithms were applied to data C, and Jaccard index eq.(14) was evaluated for segmentation result of each tested sequence frame by frame. Graphs of the Jaccard index are shown in Figure 10. The results show similar tendency as for data A and B. SRG algorithm yields quite unstable results. The original SMC based algorithm shows high values of J for sequences No.1-3, but fluctuations are observed. The results of original algorithm for sequences No.4-6 shows lower values of J. Road in this images have multiple traffic lines, and due to the

issue discussed in subsection 2.1, other traffic lines are not segmented properly. The proposed algorithm shows lower magnitude of fluctuations and overall shows high values of J. Several decreases of J can be observed in sequence No.5. Let us now look more closely on segmentation results fore such frames. Figure 11



(a) SRG Based (b) SMC Based Original (c) SMC Based Proposed

Figure 10: Jaccard Indexes for Real Image Sequence

shows result for frame 125 of sequence No.5. This is a situation immediately after another car has passed an opposite traffic line. Since samples disappear from road parts occluded by passing car, these parts are not segmented, and a few iterations of the algorithm are needed to spread samples over the previously occluded road parts. Due to this the Jaccard index decreases for few frames, because the segmented region differs from its ground truth.

The evaluation results for data A, B, C described above shows, that the proposed algorithm pose an effective solution to the issues discussed in subsection 2.1.

Figure 11: Frame 125 of sequence No.5

4. CONCLUSION AND FURTHER WORK

This paper has presented a road segmentation algorithm based on sequential Monte-Carlo estimation. The proposed algorithm is an extension of our previous work on topic of road region segmentation, and the proposed algorithm solves issues remained in original SMC based segmentation algorithm.

The key idea of the proposed algorithm lies in improvement of sampling method and consequent calculation method of importance weights. Drawing samples according to distribution of importance weights, which was adopted in the original algorithm, was not optimal for road region segmentation, and more general approach based on principle of importance sampling has been introduced. Particularly, uniform sampling from region segmented at previous time step was introduced to generate new sample set, and assignment between subsequent samples was introduced to evaluate transition probability density function required in weight update.

The proposed algorithm was evaluated on three types of image data, and evaluation results shows superiority of the proposed algorithm over the original one, as well as over the algorithm based on seeded region growing.

In the proposed algorithm, RGB colour components were used as features, however the proposed algorithm can be easily adapted to different features too. Evaluation of the proposed algorithm for features which are more resistant to illumination changes, and adaptation of the algorithm for road images with hard-cast shadows is a main direction for future work.

REFERENCES

[1] Y. Wang, D. Shen, and E. Teoh, "Lane detection using catmullrom spline," in IEEE International Conference on Intelligent Vehicles, 1998, pp. 51–57.

[2] B. Ma, S. Lakshmanan, and A. O. Hero, "Pavement boundary detection via circular shape models," in Proceedings of the IEEE Intelligent Vehicles Symposium 2000, 2000, pp. 644–649.

[3] Y. Wang, E. Teoh, and D. Shen, "Lane detection and tracking using b-snake," ImageVision and Computing , vol. 22, no. 4, pp. 269–280, Apr. 2004.

[4] H. Xu, X. Wang, H. Huang, K. Wu, and Q. Fang, "A fast and stable lane detection method based on b-spline curve," in IEEE 10th International Conference on Computer Aided Industrial Design Conceptual Design. Ieee, 2009, pp. 1036–1040.

[5] X. Miao, S. Li, and H. Shen, "On-board lane detection system for intelligent vehicle based on monocular vision," International Journal on Smart Sensing and Intelligent Systems , vol. 5, no. 4, pp. 957–972, Dec. 2012

[6] M. Foedisch and A. Takeuchi, "Adaptive real-time road detection using neural networks," in Proc. 7th Int. Conf. on Intelligent Transportation Systems, Washington D.C , 2004.

[7] S. Zhou, J. Gong, G. Xiong, H. Chen, and K. Iagnemma, "Road detection using support vector machine based on online learning and evaluation," in Intelligent Vehicles Symposium 2010, Jun. 2010, pp. 256–261.

[8] E. Shang, X. An, L. Ye, M. Shi, and H. Xue, "Unstructured road detection based on hybrid features," in Proceedings of the 2012 2nd International Conference on Computer and Information Applications (ICCIA 2012), Dec. 2012, pp. 926–929.

[9] J. Crisman and C. Thorpe, "Scarf: A color vision system that tracks roads and intersections," IEEE Trans. on Robotics and Automation, vol. 9, no. 1, pp. 49 – 58, February 1993.

[10] O. Ramstrom and H. Christensen, "A method for following of unmarked roads," in Proceedings of Intelligent Vehicles Symposium 2005. IEEE, 2005, pp. 650–655.

[11] S. K. O. Changbeom, S. Jongin, "Illumination robust road detection using geometric information," in 15th International IEEE Conference on Intelligent Transportation Systems (ITSC2012), 2012, pp. 1566–1571.

[12] J. Alvarez-Mozos, A. Lopez, and R. Baldrich, "Illuminant-invariant model-based road segmentation," in IEEE Intelligent Vehicles Symposium 2008 , Eindhoven, Netherlands, Jun. 2008, pp. 1175 –1180.
[13] Z. Prochazka, "Road region segmentation based on sequential monte-carlo estimation," in Proceedings of ICARCV 2008 , Dec. 2008, pp. 1305–1310.
[14] Z.Prochazka, "Pathway estimation for vision based road following suitable for unstructured roads," in Proceedings of ICARCV 2012 , Dec. 2012, pp. 1374–1379.
[15] M. A. Isard, "Visual motion analysis by probabilistic propagation of conditional density," Ph.D. dissertation, University of Oxford, Sep 1998.
[16] Z. Prochazka, "Road tracking method suitable for both unstructured and structured roads," International Journal of Advanced Robotic Systems , vol. Vol. 10, no. No. 158, pp. 1–10, 2013.
[17] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," Statistics and Computing, vol. 10, no. 3, pp. 197–208, 2000.

AUTHORS

Zdeněk Procházka received M.Sc degree in radio-electronics in 1994 from the Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic. In 2000 received Ph.D degree from University of Electro-Communications in Tokyo, Japan. From 2007 associated professor of National Institute of Technology, Oita College in Oita, Japan.

INTENTIONAL BLANK

EFFECT OF GRID-ADAPTIVE INTERPOLATION OVER DEPTH IMAGES

Arbaaz Singh

Department of Computer science & Engineering,
Indian Institute of Technology, Ropar, Punjab, India
arbaazs@iitrpr.ac.in

ABSTRACT

A suitable interpolation method is essential to keep the noise level minimum along with the time-delay. In recent years, many different interpolation filters have been developed for instance H.264-6 tap filter, and AVS- 4 tap filter. This work demonstrates the effects of a four-tap low-pass tap filter (Grid-adaptive filter) on a hole-filled depth image. This paper provides (i) a general form of uniform interpolations for both integer and sub-pixel locations in terms of the sampling interval and filter length, and (ii) compares the effect of different finite impulse response filters on a depth-image. Furthermore, the author proposed and investigated an integrated Grid-adaptive filter, that implement hole-filling and interpolation concurrently, causes reduction in time-delay noticeably along with high PSNR .

KEYWORDS

Depth Images, Hole filling, Interpolation, Interpolation filter

1. INTRODUCTION

In 3-D computer graphics, a depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint [1]. Once the original image and depth image is given, 3-D can be synthesized by mapping pixel coordinates one by one according to its depth value. It is the next emerging revolution after the high definition video and is the key technology in advanced three dimensional television systems (3-D TV) and free-view television systems [2-4]. A new member of 3-D sensor family, kinect has drawn great attention of researchers in the field of 3-D computer vision for its advantage of consumer price and real time nature. Based on a structured light technique, Kinect is able to generate depth and colour images at a speed of about 30 fps [5]. However, limited by depth measuring principle and object surface properties, the depth image captured by the Kinect contains missing data as well as noise. These areas of missing data are known as holes. Holes appear due to sharp horizontal changes in depth image, thus the location and size of holes differ from frame to frame. Several attempts are made to remove the noise and filling of holes with the correct data to make it suitable for different applications by means of bilateral and median-filters [6]. Apart from hole-filling, image interpolation as well occurs in all digital pictures at several stages [7]. Interpolation is the process of determining the values of a function at positions lying between its samples. It achieves this process by fitting a continuous function through the discrete input samples. This permits input values to be evaluated at arbitrary positions excluding those defined at the sample points. Interpolation is required to produce a larger image than the one captured and finds an imperative

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 185–189, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41214

consign in transmission of 3-D images. 3-D images have been used in robotic guidance, product profiling and object tracking, in battle preparation, medical diagnosis and many more [8]. These all applications require both hole-filling and interpolation to provide a preferred output. A suitable interpolation method is essential to keep the noise level minimum along with the time-delay. The standard/conventional procedure to interpolate a depth image is to first fill the holes and then apply the interpolation filter that leads to low *PSNR* and a great time-complexity. In recent years, many different interpolation filters have been developed for instance H.264-6 Tap filter, AVS- 4 tap filter and so on [9-10].

The standard/conventional procedure to interpolate a depth image is to first fill the holes and then apply the interpolation filter that leads to low *PSNR* and a great time-complexity. Recently, a texture-adaptive hole-filling algorithm is proposed for post-processing of rendered image on 3D video to save the computational cost [11]. The algorithm first determines the type of holes, and then fills the missed pixels in raster-order depending upon hole types and texture gradient of neighbors with simple data operation, which benefits for fast processing. Further, the quality of virtual view images from rendering is demonstrated by establishing a connection of pixel coordinate warping between reference image and virtual image in the rendering process, to make a quick decision on the position of hole region and its edge [12]. Subsequently, by outward expanding the edge of holes, warping error pixels are covered. Then, hole-filling is through using use mean filter and the image restoration method after eliminating the false contour by image synthesis. Nonlinear filters called Spline Adaptive Filters (SAFs), implementing the linear part of the Wiener architecture with an IIR filter instead of an FIR one are also come into picture to improve the PSNR and computational delay [13]. This paper investigates a four-tap Grid-adaptive filter on a hole-filled depth image that provides uniform interpolations for both integer and sub-pixel locations in terms of the sampling interval and filter length. Further, it compares with other different finite impulse response filters and investigates the integrated grid adaptive filter to reduce the time delay.

2. THEORETICAL & EXPERIMENTAL ANALYSIS

The depth image *I* is a 2D grid of K_V x K_H pixels. The pixels denote the distance of the objects in real scene according to its image co-ordinates [4]. The holes are areas in the image that have invalid (missing) data. Each pixel has 8 neighbouring pixels that share a face or a vertex with the centre pixel. Our goal is to go through the image and fill up the holes in the image. Let each pixel in the image be denoted as *I (v, h)*, where *v, h* can be any integer value between *0* and K_V, K_H respectively. If *I (v, h) = 0*, then that pixel is a part of a hole. A *3 x 3* gaussian weighted averaging filter is used to fill the holes. The filter takes the weighted average of the depth values of its neighbours and replaces the hole with the obtained averaged value. The weights are decided such that the neighbour pixels, that are holes themselves, are ignored while the non-zero valued pixels are used to find the new value of the hole-pixel.

Let the vertical and horizontal distances between the two nearest known pixels be T_v and T_h, respectively. Subsequently, the aim is to insert and interpolate N_v and N_h pixels within the intervals of T_v and T_h respectively. So, after the interpolation, we will have a sum of $[K_v(1+N_v)-N_v]\times[K_h(1+N_h)-N_h]$ pixels and the sampling intervals for vertical and horizontal directions will be changed to $D_v = T_v/(1+N_v)$ and $D_h = T_h/(1+N_h)$ respectively. For example, Figure 1 shows the case of $N_v = N_h = 1$ with $D_v = T_v/2$ and $D_h = T_h/2$.

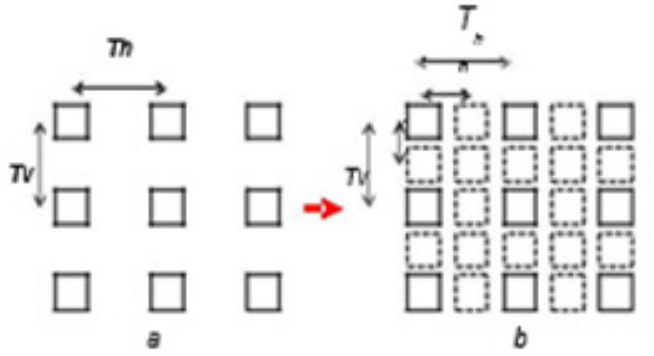


Figure 1. Uniformly spaced interpolation (a) before interpolation with known (bold square) pixels, (b) After Interpolation for missing (dotted square) pixels

Therefore, the general formula for 1-D interpolation filter will be

$$y(KT + n\Delta) = \sum_{m=M+1}^{M} y(KT + mT)P(mT - n\Delta) \quad \ldots (1)$$

Where, n denotes the pixel location to be interpolated within two reference samples i.e. *(KT)* and *(KT + T)*. So, we need to evaluate equation (1) for all *(1-N)*. The expression (1) covers both fractional and integral pixel interpolations. For the fractional locations, the value of Δ is kept less than 1 while for integral locations the value of Δ is always greater than 1. The proposed-4 tap filter is obtained using the Langrage interpolation, which is used to generalise the linear interpolation by approximating the *sinc* function [10]. The Lagrange interpolation kernel is an L^{th} order polynomial function determined by *L+1* values in the following function,

$$P(t + mT - n\Delta) = \frac{Q_{mT-n\Delta}(t)}{Q_{mT-n\Delta}(mT - n\Delta)} \quad \ldots (2)$$

Where,

$$Q_{mT-n\Delta}(t) = \prod_{k} (t - (kt - n\Delta)) \Big/ (t - (mt - n\Delta)) \quad \ldots (3)$$

For any sampling grid layout and scale, the filter coefficients can be calculated by fitting (3) to the grid. The *PSNR* is measured for different filters such as the linear averaging filter, H.264-6 Tap filter with coefficients *(1,-5, 20, 20, -5, 1) / 32*, the AVS- 4 Tap filter with coefficients *(-1, 5, 5,-1) / 8* and the proposed-4 tap grid adaptive filter with coefficients *(-1, 9, 9, -1) / 16* to recommend the best one. All test images are taken from the Middlebury Database *(2006)* and are expanded as the same sampling layout and scale of the proposed 4-Tap filter i.e. doubling the number of rows and columns with *T = 2, N = 1, M= 2, and Δ = 1*). For evaluating the performance the simulated filters, all images are expanded at the same sampling layout and scale of the filters (i.e. doubling the number of rows and columns with $T = 2, N = 1, M = 2$, and $D = 1$). For computation, the images are reduced to half before interpolation so that the size of image remains same after hole-filling and interpolation. The *PSNR* is calculated between a perfect image and its noisy approximation and can be easily defined via mean squared error

(MSE). For a given noise-free $m \times n$ monochrome image $'I'$ and its noisy approximation $'K'$, MSE is defined as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i, j) - K(i, j)^2 \quad \ldots (4)$$

and $PSNR$ is measured as

$$PSNR = 10 \log_{10} \left(\frac{I^2{}_{MAX}}{MSE} \right) = PSNR = 20 \log_{10} \left(\frac{I_{MAX}}{MSE} \right) \quad \ldots (5)$$

As shown in the Table 1, although the filter length is shorter than H.264-6, the proposed 4-Tap filter yields the highest average PSNR. The obtained results show that the proposed-4 filter with the coefficients *(-1, 9, 9, -1) / 16* has almost equal influence to the H.264 filter in terms of $PSNR$ evaluation. Consequently, it is suggested to recommend H.264-6 filter to produce high quality images. Otherwise, the proposed-4 filter is a better option that provides high quality images along with minimal time processing delays as it is used four level tapping. Further, besides the $PSNR$ values, the time-complexity of the proposed integrated adaptive filter is also evaluated. Conventionally, RGBD images are enhanced by implementing interpolation and hole-filling algorithm independently but it leads to a large amount of time-complexity along with low $PSNR$. In this work, the author proposed an integrated adaptive filter that implements hole-filling and interpolation concurrently. The time-complexity of the conventional filter is deliberated as $O\left(9 * n^2\right)$ though the proposed integrated filter has a time-complexity of $O\left(8 * n^2\right)$. Accordingly, it is obvious that a reduction of $O\left(n^2\right)$ is reported through the projected filtering scheme.

Table 1. Measured PSNR using different Filters

Image	Measured PSNR			
	Linear Average Filter	**AVS- 4 Tap Filter**	**H.264- 6 Tap Filter**	**Grid Adaptive-4 Tap Filter**
Aloe	39.9108	40.0318	40.0684	40.0846
Baby 2	47.109	47.2672	47.2149	47.2845
Baby 3	47.1809	47.4044	47.2218	47.4055
Bowling 2	44.2	44.4352	44.3577	44.4295
Cloth	47.6444	47.7766	47.7397	47.8162
Cloth 3	49.6144	49.7569	49.6258	49.7836
Computer	42.7804	43.0139	42.0716	42.9928
Flowerpots	45.0702	45.2199	45.0892	45.2617
Lampshade 2	44.0775	44.3312	44.1833	44.3026
Rocks	51.6279	51.8097	51.6726	51.8243

3. CONCLUSIONS

To demonstrate the power of grid adaptive filter, the PSNR is measured for different filters to determine the best one. In our experiments, it has been shown that the grid adaptive four-tap filter yields the highest average $PSNR$ values (almost same as that of the six-tap filter). Moreover, it offers minimum time delays than that of six-tap filter on integrating hole-filling and interpolation

in tandem. Accordingly, it is suggested to use the proposed integrated Grid adaptive filter for the enhancement of depth images.

ACKNOWLEDGEMENT

I thank to Professor Chee Sun Won, Department of Electronics and Electrical Engineering, Dongguk University, Seoul, South Korea for his kind and valuable guidance for carrying out this work. I also thank to the Department of Electronics and Electrical Engineering, Dongguk University, Seoul, South Korea for providing lab facility and technical support.

REFERENCES

[1] Depth Map, Wikipedia accessed on 17th July, 2014, http://en.wikipedia.org/wiki/Depth_map
[2] Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, (2007) "Multi-view Imaging and 3DTV", IEEE Signal Processing Magazine, Vol. 24, No. 6, pp 10-21.
[3] T. Fujii, and M. Tanimoto, (2002) "Free viewpoint TV system based on ray-space representation", in proceeding of Three Dimensional TV, Video and Display, 175, SPIE 4864, 1st November, 2002, doi:10.1117/12.454905
[4] S. L. Forman, and L. A. Steen, "Case study: Image Processing", http://www.stolaf.edu/people/steen/Projects/ ATE/imp.html.
[5] Li Chen, Hui Lin, and Shutao Li, (2012) "Depth image enhancement for Kinect using region growing and bilateral filter", 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba, Japan, pp 3070 - 3073.
[6] S. Matyunin, D. Vatolin, Y. Berdnikov, and M. Smirnov, (2011) "Temporal filtering for depth maps generated by Kinect depth camera", 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), pp 1-4, Doi: 10.1109/3DTV.2011.5877202
[7] Digital Image Interpolation, http://www.cambridgeincolour.com/tutorials/ image-interpolation.htm
[8] Andrew Wilson, "3D imaging systems target multiple applications", http://www.vision-systems.com/articles/print/volume-18/issue-9/features/3d-imaging-systems-target-multiple-applications.html
[9] Lehmann, T.M., Gonner, C., and Spitzer, K., (1999) 'Survey: interpolation methods in medical image processing', IEEE Trans. Med. Imag., Vol. 18, No. 11, 1999, pp 1049–1075.
[10] Chee Sun Won, (2013) "Grid Adaptive Interpolation Filter", Electronic Letters, Vol. 49, No. 3, pp 181-182, Doi: 10.1049/el.2012.2481
[11] Linwei Zhu, Mei Yu, Gangyi Jiang, Xiangying Mao, Songyin Fu, Ting Luo, "A New Virtual View Rendering Method based on Depth Map for 3DTV3DTV; virtual view rendering; depth; false contour; holes filling; image restoration", Procedia Engineering, Volume15, pp.1115-1119, 2011
[12] Michele Scarpiniti, Danilo Comminiello, Raffaele Parisi, Aurelio Uncini," Nonlinear system identification using IIR Spline Adaptive Filters", Signal Processing, Volume 108, pp.30-35, March 2015
[13] Chee Sun Won, "Grid Adaptive Interpolation Filter", Electronic Letters, Volume 49, Issue 3, pp. 181-182, 31st January 2013, doi: 10.1049/el.2012.2481

AUTHOR

I, Arbaaz Singh am currently studying in the field of Image Processing in the department of Computer Science & Engineering at Indian Institute of Technology, Ropar, Punjab, India. I worked as a research Intern at Dongguk University under the kind guidance of Professor Chee Sun Won in the field of Depth Images. My areas of interest are Depth image processing and designing issues of Interpolation filters.

INTENTIONAL BLANK

EVENT DETECTION IN TWITTER USING TEXT AND IMAGE FUSION

Samar Alqhtani, SuhuaiLuo and Brian Regan

School of Design, Communication and IT, the University of Newcastle,
Callaghan NSW 2308, Australia
Samar.alqhtani@uon.edu.au, suhuai.luo@newcastle.edu.au, brian.regan@newcastle.edu.au

ABSTRACT

In this paper, we describe an accurate and effective event detection method to detect events from Twitter stream. It detects events using visual information as well as textual information to improve the performance of the mining. It monitors Twitter stream to pick up tweets having texts and photos and stores them into database. Then it applies mining algorithm to detect the event. Firstly, it detects event based on text only by using the feature of the bag-of-words which is calculated using the term frequency-inverse document frequency (TF-IDF) method. Secondly, it detects the event based on image only by using visual features including histogram of oriented gradients (HOG) descriptors, grey-level co-occurrence matrix (GLCM), and color histogram. K nearest neighbours (Knn) classification is used in the detection. Finally, the final decision of the event detection is made based on the reliabilities of text only detection and image only detection. The experiment result showed that the proposed method achieved high accuracy of 0.93, comparing with 0.89 with texts only, and 0.86 with images only.

KEYWORDS

Imageprocessing, Multimedia, Data Mining, Event Detection

1. INTRODUCTION

Today, the world has greatly changed, including the way people communicate. One of the most recent phenomenons has been the social media collections that are available over the Internet. Social media is simply defined as application and websites that allow the users to create and share information or content, as well as to participate in other activities such as social networking[1]. They allow people to interact, exchange information concerning their lives such as uploading photos of events and current issues going on in their lives. Today, it is not only used by people for personal purposes only, but also by organizations for corporate issues.

The growth of social media over the last one decade has been tremendous with numbers of people joining doubling almost on a daily basis. This growth has brought about the need for scalable, robust and effective techniques of managing, as well as indexing them. Anything going on in the world is shared and communicated through the internet, especially in the social media. The social media offers people the chance to interact, comment on events, and send instant messages all over the globe without geographical barriers. Some of them include Facebook, YouTube and Twitter amongst others. The social media platforms have opened many research opportunities because of the amount of information they possess. This information can be used for many purposes including things such as prediction and detection of events and even as warning systems. Events are one of the most important indications of people's memories. They are a natural way through

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 191–198, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41215

which people refer to any observable occurrences that bring people together in the same places and time to undertake similar activities [2]. They are quite useful in making sense of the world around us, as well as in helping people recall experiences they go through. Social events on the other hand refer to those events that are attended by people and presented in the multimedia content that is shared through online websites. Such events can include disasters, concerts, sporting events, public celebrations and protests amongst others.

This paper is about detecting events based on information collected from social media. Specifically, we shall use Twitter, which is one of the social media that have fast emerged over the last few years. Many people use Twitter for reporting events as they happen in the real world [3]. This social media currently has over 500 million registered accounts all over the world that generate about 340 million messages daily. Most of these messages contain personal updates, opinion concerning current issues, moods, genera life observations and events amongst others. The readily available wide range of data from Twitter offers an ideal source for mining information for this research. There have been proposals concerning event mining that make use of the Tweet texts. However, none of them has proposed used of images in the data mining, which makes the analysis solely textual. In this paper, we seek to use both textual and visual mining to detect events in order to improve the performance of the event detection considering the retrieved results will depend on the amount of information collected [4].

The paper aims to develop an accurate and effective detection method to detect events from the Twitter stream. We monitor the Twitter stream in order to pick up texts that have photos, which are then stored in a database. It is followed by an extraction of features in both text and photos to be applied in the mining stage. It uses “bag of words” as the features of the text which will be collected using the Term Frequency-Inverse Document Frequency method (TF-IDF) [5]. For visual features, it uses Histogram of Oriented Gradients (HOG) descriptors for object detection, Grey-Level Co-occurrence Matrix (GLCM) for texture description, and color histogram[6-8]. The K Nearest Neighbours (Knn) classification with Euclidean distance algorithm will be applied to mine the data and get the accuracy measure[9].

The success of the proposal will result in breakthrough in event detection based in social multimedia data, and make the mining result more effective and accurate. This paper is structured as follow: in the next section some researches and progress in the area of multimedia event detection in social media are presented, followed by the proposed method in the third section. In the last section the experiments and evaluation of event detection are described.

2. MULTIMEDIA EVENT DETECTION IN SOCIAL MEDIA

The advance of systems, in particular the web-based social systems, has heightened in recent years in an exponential manner. Recently, academics and researchers started to examine a range of data mining techniques to assist experts enhance social media [10]. These techniques permit experts to discern novel information derived from users’ application data. Lately, a range of community services as well as web-based sharing like YouTube and Flickr have made a massive and hastily mounting amount of multimedia content accessible online. Content uploaded by partakers in these vast content pools is escorted by wide-ranging forms of metadata, like descriptive textual data or social network information.

Twitter and social media trends have drastically changed in the recent past with millions of users going to the platform to chat, exchange ideas or share stories[11]. As a result, this platform has formed a rich place for news, events and information mining. However, due to the huge burst in information, data mining in Twitter is a complicated venture that requires a lot of skills and information on important ways of undertaking data mining. Twitter and social media sites have

traffic overflows which are multiple and huge in terms of the frequency [12]. For instance twitter receives over 80 million tweets a day and this leads to billions of tweets per month. As a result, event prediction and detection requires the use of complex algorithms which go through the text and images in keyword matching process[13]. One of the requisite skills includes extraction features and algorithms that could be deployed in mining data such text and data. Several data mining tools and algorithms have been developed with the capabilities and purpose of analysing data and text. Researchers and other people have come up with techniques of mining data from social media with the use of different types of algorithms[14]. For instance, through the use of mining tools such as RData Mining tool, we can target some key words to mine within events and other forms of data from twitter streams. There are several techniques that could be used in the process of mining text and multimedia data in social media channels [15]. One of the important uses of data mining within social media is on event detection in twitter or social media channels. Event detection within social media through the use of different data mining techniques and algorithms is common and growing within the social media sphere [16]. Techniques such as mining events through geo-tagged events and geo-tweet photos have been utilized in regions such as Japan and Singapore to identify events such as Typhoons and floods. These techniques have been successfully in finding information on different events. These mining techniques make use of keywords to within bursts of Twitter streams for matching identities [17]. As a result, these tweets are grouped into certain databases where they are analysed. The processes of mining involved searching for keywords with emphasis on event detection with focus on words those are frequent. Then these event keywords will be unified while geo-tweet photos which correspond with the keywords will be clustered and grouped together. Each of these photos will be matched against these events and shown on the map. In the process of event detection it is imperative to look into variable factors such as distinct languages and locations. This approach will address the gap found in the process and tasks which require quick event detection in Twitter and social media circles [18].

Twitter is one the social media sites that have tremendous traffic overflows which are multiple and huge in terms of the frequency. For instance twitter receives over 80 million tweets a day and this leads to billions of tweets per month. As a result, event detection requires the use of complex algorithms which go through the text in keyword matching process [17]. For text mining in event detection by using Twitter data, there are different way to detect event like using part of speech technique, Hidden Markov Model (HMM), and Term Frequency and Inverse Document Frequency (TF_IDF).

As the saying goes a picture is worth a thousand words. Nowadays social media users find it much more convenient and enjoyable than ever before to express their opinions by posting pictures, attaching video clips rather than just typing a message. Mobile social network application developers also introduce features to allow users to take pictures and then upload them through a simple click. Compared with text information, multimedia contents are more eye-catching and entertaining. Social media sites are used in posting multimedia data such as photos, videos and other content which allow information sharing such as events. In the process of event detection we have to make use of algorithms and techniques that allow searching, extraction and storing of multimedia data from social media streams. Due to the huge volume of content such as videos, images and other content, we have to utilize techniques with emphasis on content-basis image retrieval algorithms [14].

This paper will look into data mining for the purpose of event detection through the use of two algorithms in the process of extraction. These methods will utilize different algorithms in extracting text and photo streams from twitter. The failure of having an accurate event detection method in the process of social media or twitter mining precipitates a problem that needs a solution [19]. As a result the use and combination of Term Frequency-Inverse Document

Frequency method (TF-IDF) and for image content, Histogram of Oriented Gradients (HOG) descriptors for object detection, Grey-Level Co-occurrence Matrix (GLCM) for texture description, and color histogram might yield better results in the process of event detection. These algorithms have been effective in the process of mining and obtaining information on different events.

3. THE PROPOSED METHOD

In this section, we explain the detail of each step of the proposed system. Before that, we monitor Twitter stream to pick up tweets having both text and photos, and store them into a database. Then, we detect the event in text data only, image data only, and fuse the image with the text in last method.

3.1. Text Data Mining

Text data mining is useful for research into social media because it gives researchers the ability to automatically detect event in Twitter. We use the text data to detect event in this step and our method is depicted in figure 1.

Tweet message is written in sentence in general of which the maximum number of letters is 140. To do event detection by using text data in Twitter, we filtered out tweets that contain non –Latin characters, trying to maintain a corpus of English tweets. Although we managed to remove all East Asian tweets, our corpus still contained some non–English tweets mainly in Spanish and Dutch. Lowercase all words in the tweets. Then we follow the procedure:

a) Tokenize

   Convert the string to a list of tokens based on whitespace. This process also removes punctuation marks from the text.

b) Stop word filtering

   Eliminate the words which are common and their presence does not tell us anything about the dataset, such as: the, and, for, etc.

c) Stemming filtering

   Reduce each word to its stem, removing any prefixes or suffixes.

d) Indexing

   We index the data after filtering and stemming by using TF-IDF which is a weighting scheme that weighs features in tweets based on how often the word occurs in an individual tweet compared with how often it occurs in other tweets. Then by measuring the weight for each keyword, we can decide the related event.

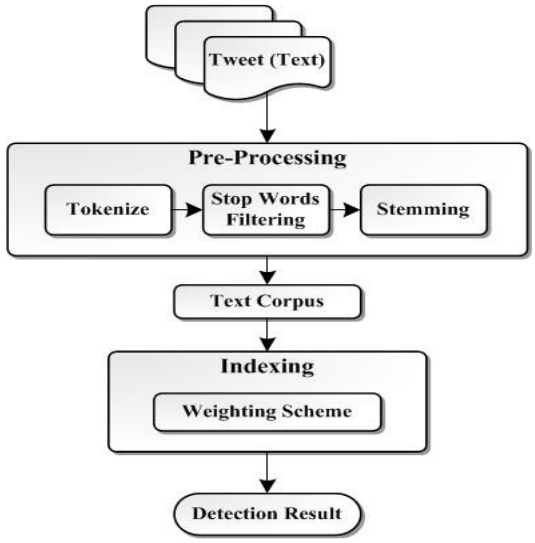


Figure 1.The block diagram of text data mining

3.2. Image Data Mining

Image mining uses three distinguishable types of feature vectors for images description to ensure the accuracy is very high at any particular case. These feature vectors are Histogram of Oriented Gradients (HOG) descriptors for object detection, Grey-Level Co-occurrence Matrix (GLCM) for texture description, and color histogram. The method is depicted in figure 2.

HOG descriptors are used in computer vision and image processing for object detection. This technique works on the occurrences of gradient orientation in localized portions of a particular image. The object appearance and the shape within an image are described by the distribution of intensity gradients. GLCM is used for purposes of texture description such as land surface or even an extensive ocean. It is defined as the distribution that is defined over an image to be the distribution of co-occurring values at a given offset values. Its main applicability is to measure the texture of surfaces. Another aspect of feature extraction is the color histogram. This aspect is used in image processing and photography and it is defined as the representation of distribution of colours of an image.

In addition to features extraction, we build a data model by using the data mining techniques. In our work, we have used K Nearest Neighbours (KNN) classification which is defined as a simple user defined algorithm or a program that stores available data cases and therefore classifies new data cases based on the principle of measure similarities such as the functions of distance. This method of classification is used in our work with Euclidian distance.

3.3. Fusion Images and Text Data

For multimedia data, we apply fusion for text and image by combining text and image features. We use our database that contains both text and photos to the proposed event mining system which consists of event keyword detection, event photo classification photo selection. The features we extract in this step are the combination of features in section 3.1, and section 3.2.

In our fusion method, if the tweet text mining score is less than a threshold, this means that the text mining is not reliable so the tweet is classified using the image only; otherwise, the tweet is classified using the text.

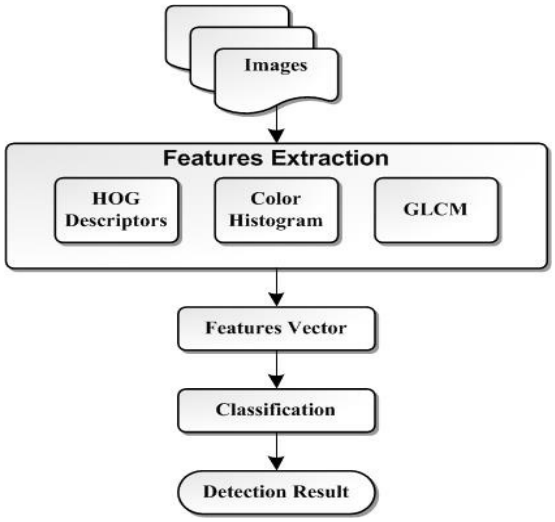


Figure 2.Theblock diagram of image data mining

4. EXPERIMENTAL RESULTS

In the experiment, we used about one million tweets which contain texts and photos posted about Napa Earthquake 2014, California, which were collected from the Twitter stream from 25August 2014 to 30 August 2014.We train our algorithms on our data. We divided the data into three equal parts. We use the earliest two thirds of the data as training and validation sets.

As results of event keyword extraction from our text data, we obtained 100 keywords related to earthquake event such as: earthquake, shock, shake, chill, Napa and others. Then, we produced a composite weight for each term. For image, we classify our data by using K Nearest Neighbours (KNN) classification into two classes depending on the event. Class 1 represents earthquake is happened, and class 2 represents earthquake is not happened. By preparing a training set, we can produce a model to classify tweets automatically into each class. The figure 3 represents the image's sample for the two classes.

Figure 3. Shows some example images for both classes where class1 meanearthquake is happened, and class 2 means earthquake is not happened

We prepare three groups of features for each tweet to detect the event as follows:

- First group features is the text features for text mining.
- Second group features is the features for image mining.
- Third group features is the fusion for text and image features.

Finally, we measure the accuracy for each mining method by apply the following equation.

$$A = \frac{TP + TN}{TP + TN + FP + FN},$$

where A represents the accuracy for the event detection method, TP, TN, FP and FN represents true positive, true negative, false positive and false negative respectively.In our classification, earthquake is happened class is a true positive.

From the experiment, we find that accuracy from event detection model for the fusion of text and image gave more accurate result and made the event detection more effective. The result is shown in Table 1.

Table 1. Result for the method's accuracy

Data Type	Text Data	Image Data	Fusion Images and Text Data
Accuracy	0.89	0.86	0.93

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed an event detection method to detect events from Twitter stream, by applying mining tool for Twitter streams that have texts and photos. It has proved that mining both visual and textual information will give accurate and effective result. In particular, we achieve better accuracy when we fuse text and image in mining algorithm for event detection application. Future work will focus on using different method of fusion for text and image features, and adding more effective features, which makes the event detection better.

ACKNOWLEDGEMENTS

The primary author and related research is sponsored by Najran University in Saudi Arabia.

REFERENCES

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, pp. 59-68, 2010.
[2] J. Mingers, "The paucity of multi method research: a review of the information systems literature," Information Systems Journal, vol. 13, pp. 233-249, 2003.
[3] G. Guthrie, "Research methodology: Basic research methods: an entry to social science research," pp. 38-50, 2010.
[4] N. Walliman, "Research theory, Research methods: the basics. ," pp. 15-28, 2011.
[5] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, pp. 513-523, 1988.

[6] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man and Cybernetics, pp. 610-621, 1973.
[7] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," presented at the Proceedings of the ninth ACM international conference on Multimedia, Ottawa, Canada, 2001.
[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE 2005, pp. 886-893.
[9] X. Wu and V. Kumar, The top ten algorithms in data mining: CRC Press, 2010.
[10] C. A. Bhatt and M. S. Kankanhalli, "Multimedia data mining: state of the art and challenges," Multimedia Tools and Applications, vol. 51, pp. 35-76, 2011.
[11] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," presented at the Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, 2010.
[12] N. Memon, J. J. Xu, D. L. Hicks, and H. Chen, Data mining for social network data vol. 12: Springer, 2010.
[13] M. A. Russell, "Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites," 2011.
[14] L. Tang and H. Liu, "Community detection and mining in social media," Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 2, pp. 1-137, 2010.
[15] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1-167, 2012.
[16] I. H. Ting, T. , Social Network Mining, Analysis, and Research Trends: A Phenomenal Analysis. Boston, MA: Cengage Learning, 2012.
[17] R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction: Cambridge University Press, 2014.
[18] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," GeoJournal, vol. 78, pp. 319-338, 2013.
[19] S. Marseken, Object Recognition and Categorization: Scale-Invariant feature transform. Savannah, GA: Lippincott Williams & Wilkins, 201

DIGITAL ENHANCEMENT OF INDIAN MANUSCRIPT, YASHODHAR CHARITRA

Sai Siddharth Kota, Raja Massand, Abhinaya Agrawal and Preety Singh

Computer Science and Engineering Department,
The LNM Institute of Information Technology, Jaipur, India

siddharthhp@gmail.com, rajamassand@gmail.com,
abc.abhi99@gmail.com, prtysingh@gmail.com

ABSTRACT

Over the years, many of our ancient manuscripts have been damaged by natural elements or intentionally erased and re-used to record other information. While manual preservation techniques are being carried out for the conservation of our ancient texts, digital image processing is an alternative for the archival storage of the invaluable text contained in them. To successfully recover text from such documents, it is important to understand the nature of the writing and materials on which they are written. Different imaging and processing techniques are needed, depending on the the condition of the manuscript. In recent years, modern imaging techniques have been applied to ancient manuscripts to recover writings that are not visible to the naked eye or not recognizable due to various factors. In this paper, we apply imaging techniques on an ancient manuscript, Yashodhar Charitra, and restore it digitally.

KEYWORDS

Manuscript Restoration, Noise Removal, Gaussian Bandpass filter, Thresholding, Image Enhancement.

1. INTRODUCTION

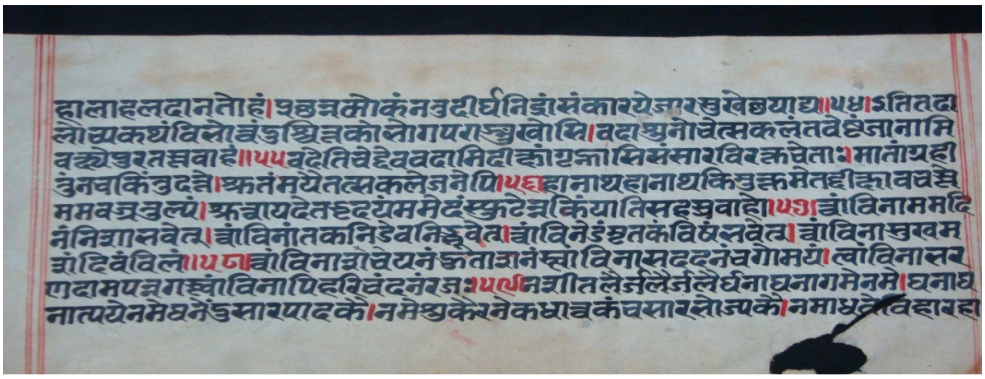
Ancient manuscripts form an integral part of our rich heritage. A number of these texts have been written on parchment. If stored under very dry conditions, parchment can last for thousands of years. For writing on them, early inks were made with carbon black suspended in water. This ink provided, and continues to provide, well-defined and high-contrast writing. Later manuscripts were written with iron gall ink, which was easier to make and harder to remove from the surface.

Natural and artificial agents have resulted in the deterioration of these manuscripts. Image processing techniques can be applied to images of these texts to restore them and store them digitally. However, for application of these techniques knowledge of the ink and parchment used is essential. Damaged parchment is often very dark, making any surviving text characters hard to read.

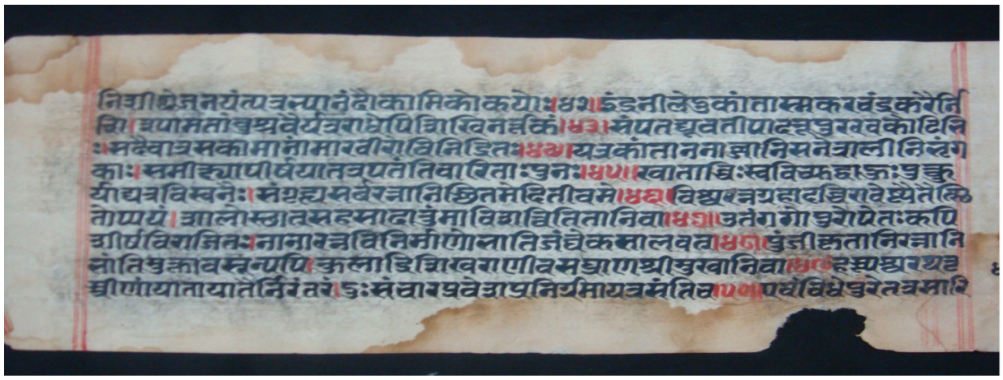
In this paper, we apply digital image processing techniques for enhancement and text restoration of an ancient manuscript, *Yashodhar Charitra*, which was written using carbon black suspended in water on parchment made up of goat or sheep skin which is very durable. Digitization of the

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 199–207, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41216

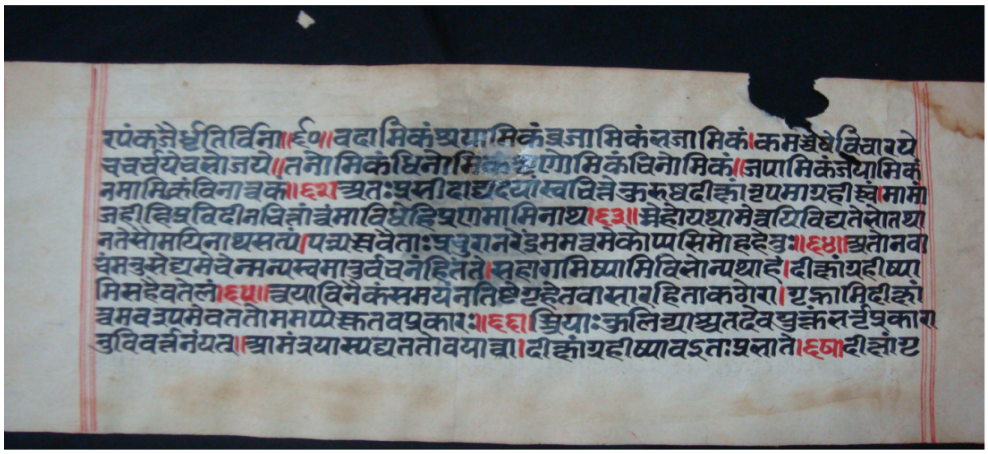
restored manuscript will ensure wider reach of the text to readers, economy of storage and safety from damage inflicted by nature.



(a) Missing text due to page tear



(b) Blotches



(c) Damage due to moisture

Figure 1: Various types of damages inflicted

2. RELATED WORK

In [**Error! Reference source not found.**], Alirezaee et al. have developed a restoration algorithm for the Pahlavi or middle-age Persian manuscripts. The algorithm uses mathematical morphology and analysis of connected components to overlapped lines, words and characters. The algorithm was tested on
200 pages of Pahlavi document. The noise and destructive effects are removed.

In [**Error! Reference source not found.**], Chaudhuri et al. have separated text from non-text doodles of poet Rabindranath Tagore in Indian manuscripts. The approach generates connected components and classifies them as text and non-text based on a comparison between the total number of pixels and the number of boundary pixels. Further separation is done based on the stroke width computed for each window.

In [**Error! Reference source not found.**], Hedjam and Cheriet proposed a data representation for text extraction from multispectral historical document images. They performed foreground pattern extraction using region-of-interest (ROI) analysis and a maximum likelihood classifier. Two new features containing spectral components are introduced.

3. PROPOSED METHODOLOGY

Manuscript restoration techniques involves various steps. First of all, we create a database of the damaged manuscript. Each page of the manuscript is photographed using a high-defintion digital camera. The manuscript is analyzed for the different damages. Each page of the manuscript is subjected to noise removal and enhancement technique. The various steps involved in our approcah are :

- Digitization of images: The manuscript was digitized using a standard digital camera.
- Noise removal: The image is filtered to remove the unwanted noise from the digitized image. This methodology uses a filtration technique which uses Gaussian Bandpass filter and the difference in colour intensities of the damaged part to the non-damaged part.
- Thresholding: It was done in order to remove the damaged segments of the image.
- Image enhacement: This helps us in improving the various aspects of an image to make it visually better.
- Text restoration: Some missing areas of the text were recovered using text restoration.

4. EXPERIMENT

4.1 Database

For experimental purpose, we obtained our database from *Digambar Jain Manuscript Conservation Centre* under *The National Mission for Manuscripts (NAMAMI)* which is an autonomous organisation under Ministry of Culture, Government of India, established to conserve Indian manuscripts and create a national resource base. We used a manuscript called *Yashodhar Charitra*, written in the year 1661, by *Gyankirti* based on the fascinating story of *Yashodhara*. It is written in *Sanskrit* with a touch of *Varhadi* dialect. It contains 69 pages. The manuscript was digitized using a digital camera. The database was analysed for various types of damages. It was observed that there were a number of factors like dark areas, blotches, degradation due to moisture and missing text due to page tear.

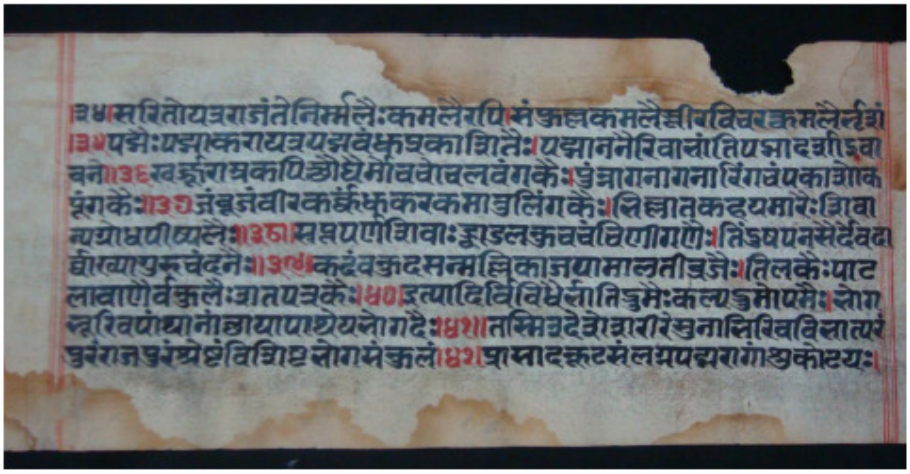
4.2 Preprocessing

Image preprocessing is very important as it can significantly increase the performance of subsequent enhancement steps. Features of digitized images along the database can vary due to different factors like luminance level, wearing out due to time etc .

4.2.1 Noise Removal

Noise removal basically acts as a tool to remove the unwanted noise from our digitized image which may have occured during image acquisition. It uses the Gaussian bandpass filter which removes noise contained in a certain range.

Gaussian bandpass filtering is done in the frequency domain. The function makes use of the simple principle that a bandpass filter can be obtained by multiplying a lowpass filter with a highpass filter where the lowpass filter has a higher cut off frequency than the high pass filter. This process is shown in Figure 2.



(a) Image before noise removal

(b) Image after noise removal

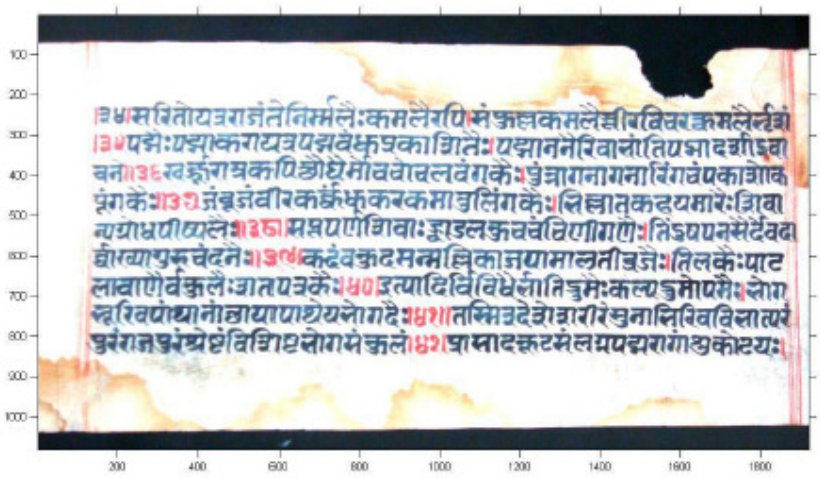
Figure 2: Noise Removal

4.2.2 Thresholding

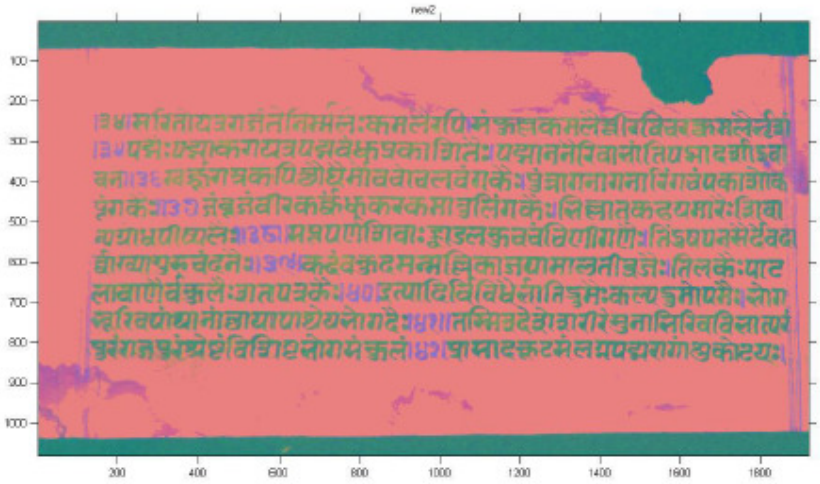
Thresholding is the simplest method of image segmentation. We used it to convert our preprocessed RGB image to various other colour formats like YCbCr, HSV, CMYK based on the type of damage inflicted.

- YCbCr was used when we had light blotches in the digital image.
- HSV was used for images degraded by moisture.
- CMYk was used for images having dark blotches.

The thresholded image is shown in Figure 3.



(a) Image before thresholding

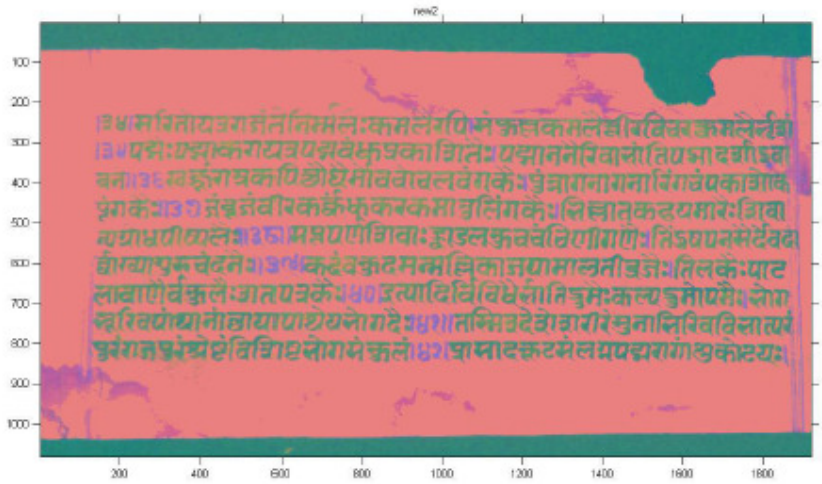


(b) Image after thresholding in YCbCr

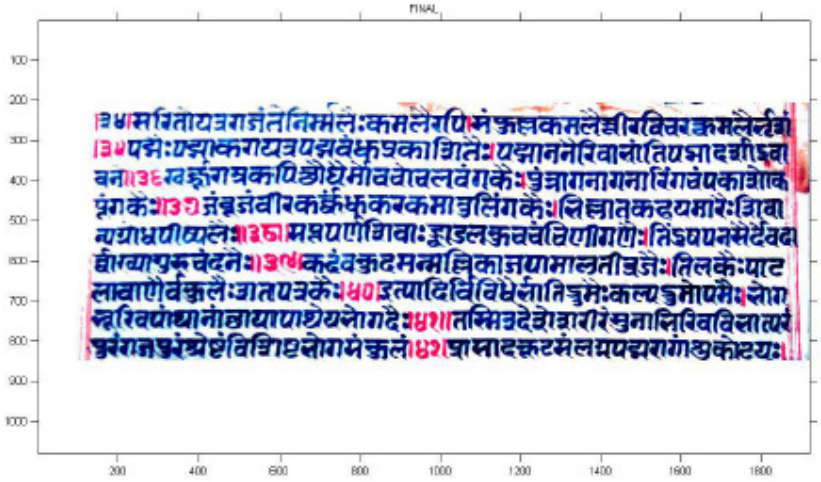
Figure 3: Thresholding

4.3 Image Enhancement

After converting the image to a specific color format based on the type of damage, it is enhaced for various aspects like brightness, contrast and sharpness to make it more visually presentable. The enhanced image is shown in Figure 4.



(a) Image before enhancement

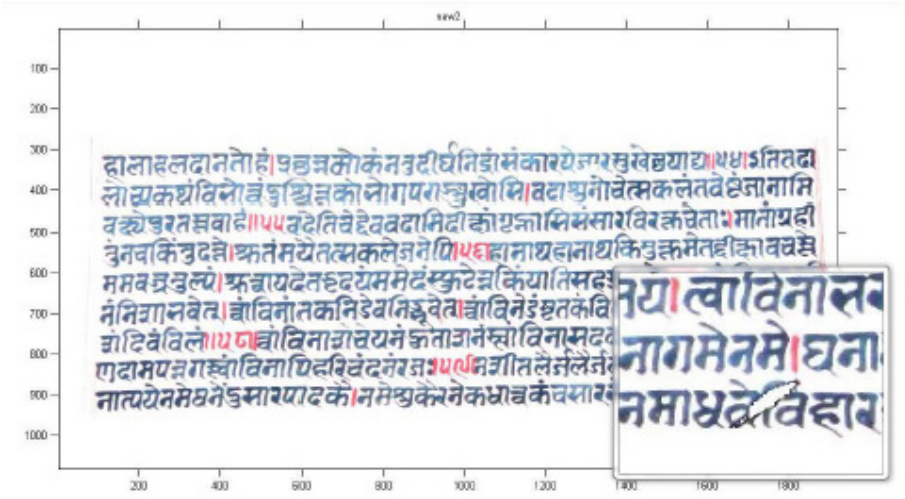


(b) Image after enhancement

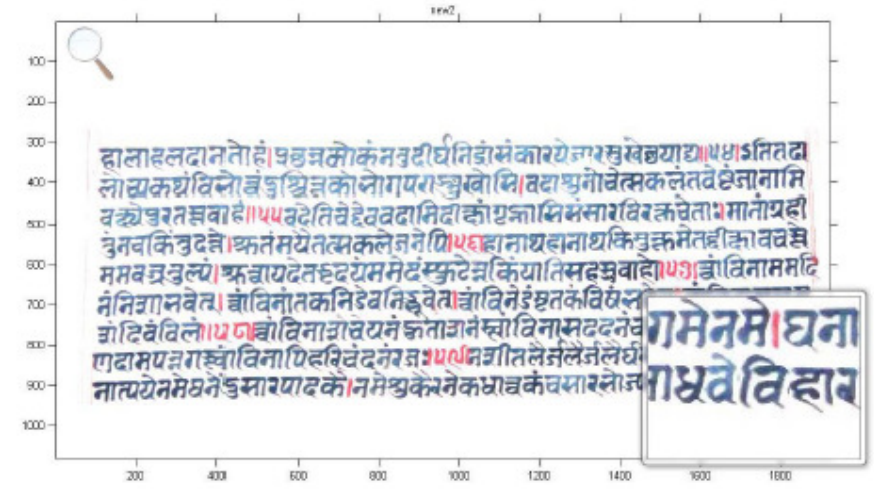
Figure 4: Image enhancement

4.4 Text Restoration

Some segments of the manuscript were severely damaged and the text in those areas was missing. For such cases, we consulted the language expert and manually replaced the missing text. Similar text is located elsewhere in the manuscript and digitally picked up and pasted where the text is missing. This can be seen more clearly in Figure 5.



(a) Image before text restoration



(b) Image after text restoration

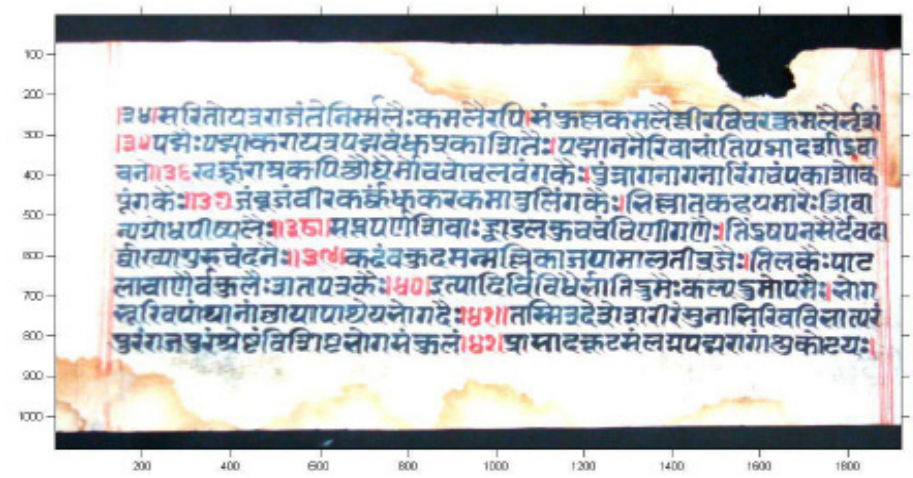
Figure 5: Text restoration

5. RESULT AND ANALYSIS

The damaged manuscript is restored using noise removal and image enhancement techniques. Text restoration for missing text is done by digitally cropping equivalent text from elsewhere and pasting it where it is missing. The results of the whole process are shown in Figure 6.

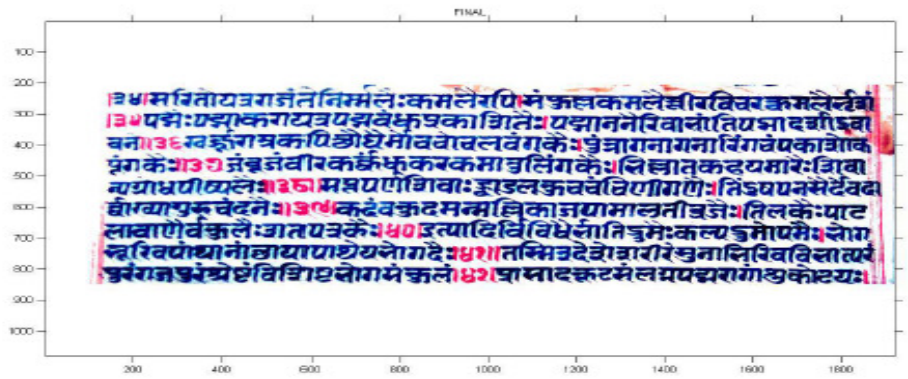
It is seen that major spoilage of the manuscript is removed. The manuscript becomes noise free. The blotches caused due to exposure to moisture, air and other pollutants are visibly removed. The text which is missing due to page tear is replaced accordingly. The end result of the restored manuscript is a clean, easy to read and defect-free digital manuscript.

(a) Image before noise removal



(b) Image after noise removal

(c) Image after thresholding



(d) Image after enhancement

Figure 6: Various stages of Manuscript restoration

6. CONCLUSION

Image processing techniques were applied for digital restoration of an ancient manuscript, *Yashodhar Charitra*, authored by *Gyankirti* . The deteriorated manuscript was inspected and treated for various defects. The noise free and enhanced manuscript can now be digitally archived for wider reach to the readers and free from effects of natural pollutants. In fututre, the techniques will be applied to other ancient manuscripts available with conservation centers.

ACKNOWLEDGEMENTS

We are thankful to Digambar Jain Manuscript Conservation Center, under National Manuscript Mission, Government of India, for their support in providing us with the manuscript.

REFERENCES

[1] Keith T. Knox, Roger L. Easton Jr. , and William Christens-Barry "Image restoration of damaged or erased manuscripts,". in 16th European Signal Processing Conference (EUSIPCO 2008), pp. 1–5, 2008.

[2] Angelika Garz, Markus Diem and Robert Sablatnig "Detecting Text Areas and Decorative Elements in Ancient Manuscripts,". in 12th International Conference on Frontiers in Handwriting Recognition, pp. 1–6, 2010.

[3] Avekash Gupta; Sunil Kumar; Rajat Gupta; Santanu ChaudhuryÎ³;Shiv Dutt Joshi "Enhancement of Old Manuscript Images," in IBM India Research Lab, pp. 1–5, 2007.

[4] Aaron Greenblatt; Karen Panetta "Restoration of Semi-Transparent Blotches in" Damaged Texts, Manuscripts, and Images Through Localized, Logarithmic Image Enhancement," in ISCCSP 2008, pp. 1–6, 2008.

[5] Umbaugh Scot E "Computer Vision and Image Processing", in Hall, NJ, 1998, ISBN 0-13-264599-8

[6] B. B. Chaudhuri, A. Saraf, A. Kumari, S. Borah and A. Goyal "Separation of text from non-text doodles of poet Rabrindranath Tagore's manuscripts", in National Conference on Computing and Communication Systems, pp. 1–5, 2012.

[7] S. Alirezaee, H. Aghaeinaia, M. Ahmadi and K. Faez "An efficient restoration algorithm for the historic middle-age Persian (Pahlavi) manuscripts", in IEEE International Conference on Systems, Man and Cybernetics, Volume 3, pp. 2114–2120, 2005.

[8] R. Hedjam and M. Cheriet "Novel data representation for text extraction from multispectral historical document image", in IEEE International Conference on Document Analysis and Recognition, pp.172–176, 2011.

INTENTIONAL BLANK

3D VISION-BASED DIETARY INSPECTION FOR THE CENTRAL KITCHEN AUTOMATION

Yue-Min Jiang[1],Ho-Hsin Lee[1], Cheng-Chang Lien[2], Chun-Feng Tai[2], Pi-Chun Chu[2]and Ting-Wei Yang[2]

[1]Industrial Technology Research Institute, SSTC, Taiwan, ROC
[2]Department of CSIE, Chung Hua University, Taiwan, ROC
[1]jongfat@itri.org.tw,[2]cclien@chu.edu.tw

ABSTRACT

This paper proposes an intelligent and automatic dietary inspection system which can be applied to the dietary inspection for the application of central kitchen automation. The diet specifically designed for the patients are required with providing personalized diet such as low sodium intake or some necessary food. Hence, the proposed system can benefit the inspection process that is often performed manually. In the proposed system, firstly, the meal box can be detected and located automatically with the vision-based method and then all the food ingredients can be identified by using the color and LBP-HF texture features. Secondly, the quantity for each of food ingredient is estimated by using the image depth information. The experimental results show that the dietary inspection accuracy can approach 80%, dietary inspection efficiency can reach1200ms, and the food quantity accuracy is about 90%. The proposed system is expected to increase the capacity of meal supply over 50% and be helpful to the dietician in the hospital for saving the time in the diet inspection process.

KEYWORDS

Dietary inspection, LBP-HF, Image depth

1. INTRODUCTION

In recent years, the food industry has been addressing the research on the food quality inspection for reducing the manpower and manual inspection error. To aim at this goal, in this study, the machine learning technologies are applied to develop the 3D vision-based inspection system[1,13] that can identify the meal categories and amount. In [2], the study indicated that the selected image features are crucial [14]to the detection of peel defects. In [3], the authors developed a vision-based method to improve the quality inspection of food products. In [4], Matsuda et al. proposed the food identification method by integrating several detectors and image features, e.g., color, gradient, texture, and SIFT features. Then, the multiple kernel learning(MKL) method is applied to identify the food quality. Yang et al. [5] proposed the pair wise local features to describe the texture distributions for eight basic food ingredients. However, the abovementioned methods do not address the quality inspection for the Chinese foods. In the Chinese food, several food ingredients are often mixed, e.g., the scrambled eggs with tomatoes, such that it is difficult to identify the food ingredients and quantity by using the conventional vision-based methods. In [8], Chen et al. proposed the diet ingredients inspection method by using the SIFT, Gabor texture, and depth camera to detect the diet ingredients. Based this method,

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 209–219, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41217

in this study, we apply the proposed the meal box detection and locating technology, LBP-HFtexture features, and depth images to construct a novel approach of the dietary inspection for the central kitchen automation. The system flowchart of the proposed 3D vision-based dietary inspection system is shown in Figure 1.

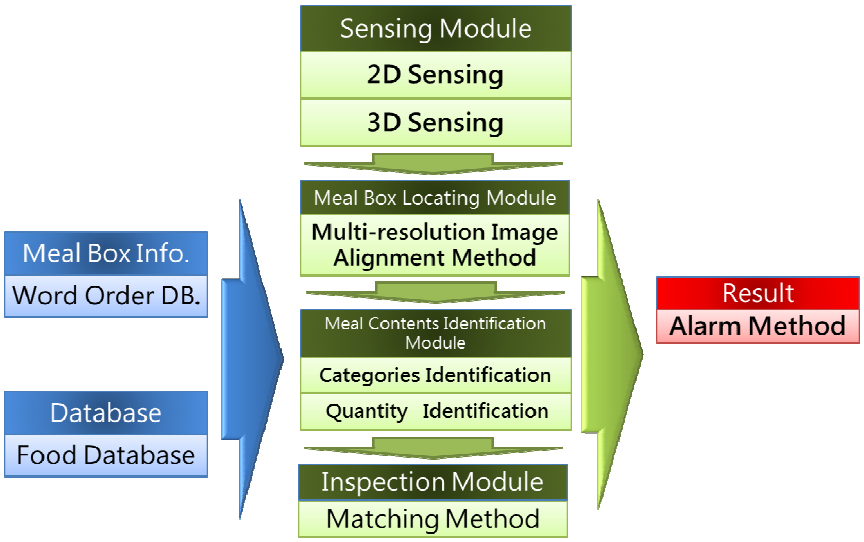


Figure 1.The system flowchart of the proposed 3D vision-based dietary inspection system.

In Fig. 1, firstly, the sensing module extracts 3D (depth) and 2D images. Secondly, the meal box locating module locates the position of the detected meal box and segment the regions for each food ingredient. Finally, the meal contents identification module identifies the food categories and amount for evaluate the food quality. The system operation procedures are described in Figure 2. The meal box is moving on the conveyor and the sensing module extracts 3D (depth) and 2D images continuously. Once the meal box is located with the meal box locating module, the food quality can be inspected with the color, texture, and depth image features. The experimental results show that the dietary inspection accuracy can approach 80%, dietary inspection efficiency can reach 1200 ms, and the food quantity accuracy is about 90%. The proposed system is expected to increase the capacity of meal supply over 50% and be helpful to the dietician in the hospital for saving the time in the diet inspection process.

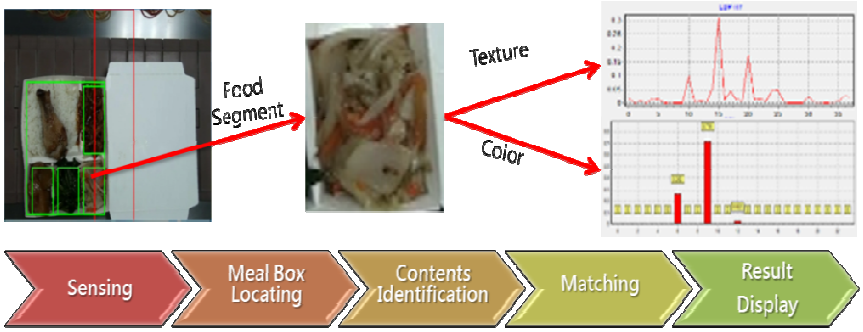


Figure 2.Thesystem operation procedures of the proposed 3D vision-based dietary inspection system.

2. AUTOMATIC MEAL BOX INSPECTION

The baseline of the system design is based on the domain knowledge of dietician. In this section, the methodologies of the meal box locating and the meal contents identification are described. In the diet content and amount identification process, the 2D/3D image features, e.g., depth, color [7] and textures[6], are used to train the dietary inspection system, and then the system can identify the diet categories and amounts. By using the novel automatically foods recognition and amount identification system, the manual operations can be reduced significantly and the accuracy and efficiency of food arrangement can be improved significantly.

2.1. Meal Box Detection and Locating with Multi-resolution Image Alignment

To develop a real-time vision-based dietary inspection system, the analyses of the video content captured from the camera are crucial to detect and locate the meal box. In Fig. 3, we can see that the meal box is moving continuously on the meal dispatch conveyor at central kitchen. Then, how to detect the meal box and locate the position of meal box in real-time is a problem. Here, we proposed a novel meal box locating method by using the multi-resolution image alignment method to match the meal box template shown in Fig. 4-(b) to the captured images from low resolution to high resolution within the region of interest (ROI) shown in Fig. 3.

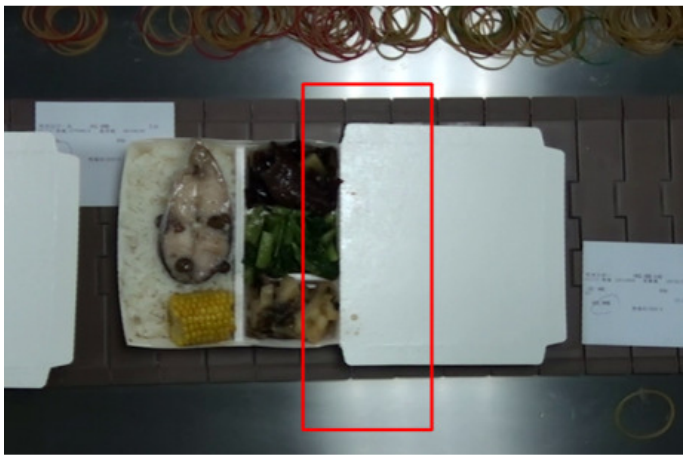
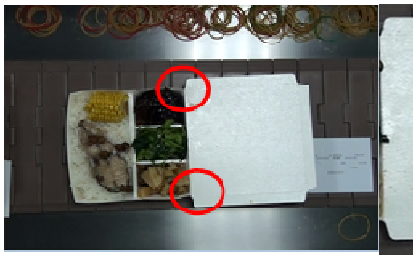


Figure3.The ROI setting in the meal box image.

Based on the careful observations, the image template of the meal box is difficult to generate because that the foods can cover the left side of meal box and no texture information exist on the right side of meal box. Hence, we extract the middle image in the meal box image shown in Fig. 4-(b) as the image template of meal box to detect and locate the position of meal box.



(a) (b)

Figure 4.Selection of image template of meal box to detect and locate the position of meal box.(a)Meal box image on the meal dispatch conveyor.(b)Image template of meal box.

The algorithm of meal box locating is described as follows.

1. Image template and meal box image are decomposed into specified multi-resolution levels (pyramid image representation) shown in Fig. 5.

2. Perform the pixel-based template matching (correlation matching) in the lower level (lower resolution). Then, some candidate regions are extracted.

3. Perform the pixel-based template matching in the higher resolution images within the neighbouring region obtained from the candidate regions in step 2.

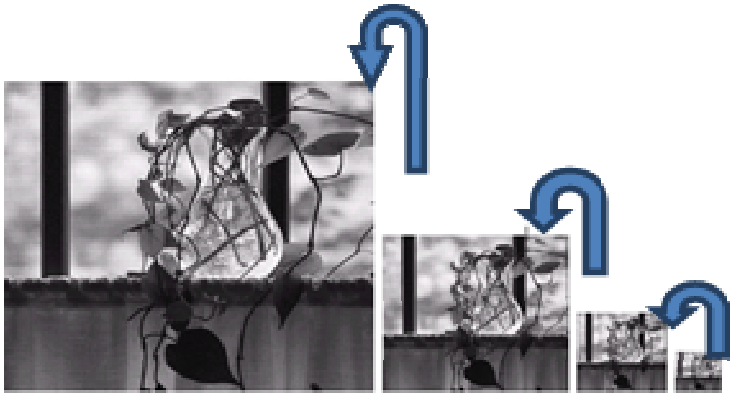


Figure 5.Image are decomposed into multi resolution levels (pyramid image representation).

To speed up the calculation efficiency, all the integer multiplication operations defined in Eq. (1) for the computation of correlation are calculated in advance and stored as a look-up table.

$$\text{GrayTable}(A, B) = \{(A - \bar{A}) \times (B - \bar{B}), 0 \leq A \leq 255, 0 \leq B \leq 255\} \quad (1)$$

where $\bar{A}$ and $\bar{B}$ are the mean values of the image A and B respectively. The system operation flowchart of the multi-resolution meal box locating module is shown in Figure 6.

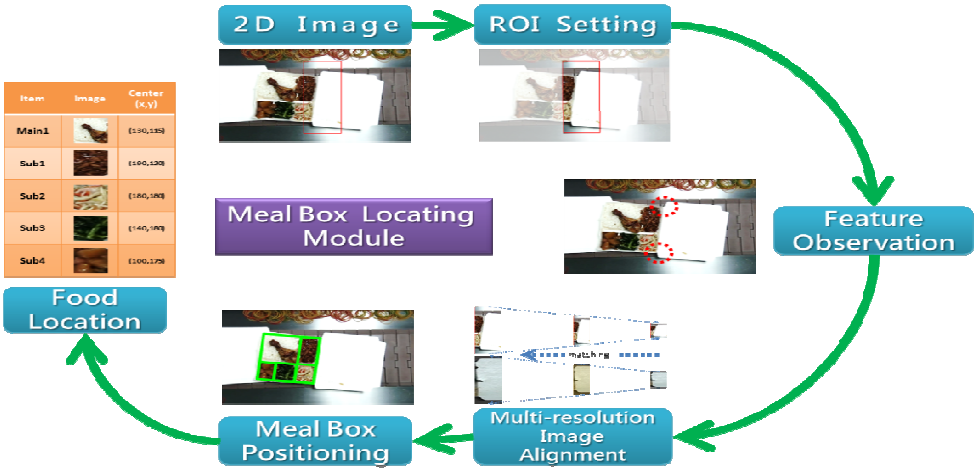


Figure6.The system operation flowchart of the multi resolution meal box locating module.

2.2. Meal Box Contents Identification

This section will describe the identification methods for the contents in the meal box, i.e., the processes in the "meal box location content identification module". The extracted features include the color distribution (polor histogram)and texture(LBP-HF) feature within the ROI. Finally, the dietary inspection is designed with the similarity measure between the online input image and the trained image features in the database. Figure 7 illustrates the flowchart of food quality inspection system.

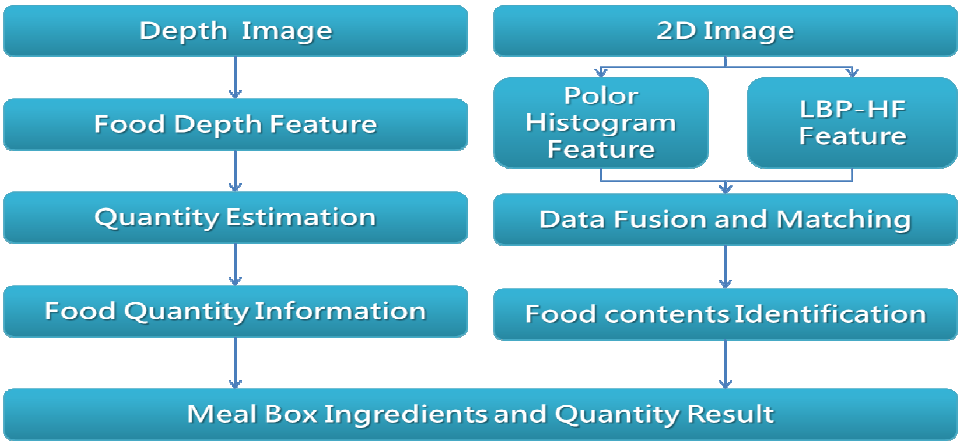


Figure7.Flowchart of food quality inspection.

2.2.1.Color Polar Histogram

Once the meal box is aligned, we can segment the regions for each food ingredient to extract the color distribution feature for identifying the food ingredient color. Here, we transfer the color space of the image of each food ingredient into YCbCr color space and use the CbCr color channels to establish the color polar histogram [7,9] with angle range from -127^{o} to 127^{o}. Figure 8 shows the polar coordinates.

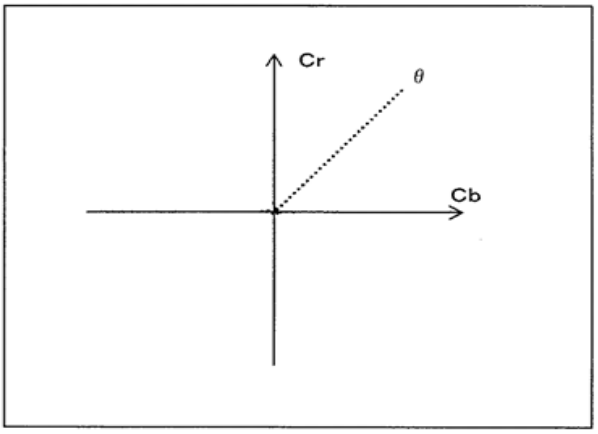


Figure 8. Polar coordinates for Cr and Cb color space.

$$h_l = \sum_{i=1}^{N} \delta[\theta(C_b, C_r) - l] \quad (2)$$

Here, the color polar histogram is represented as $H = \{h_1, \dots, h_m\}$. The value for each histogram bin can be calculated as the formula in Eq. (2), where the θ is the degree of coordinate. The extracted color polar histogram feature for the image of the food ingredient of the sample image is illustrated in Fig. 9.

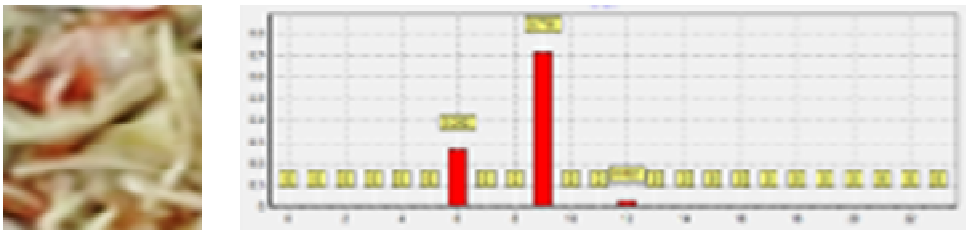


Figure 9. The extracted color polar histogram feature for the image of the food ingredient of the sample image.

2.2.2Local Binary Pattern-Histogram Fourier (LBP-HF)

Local binary pattern-histogram Fourier (LBP-HF)[10] is based on the LBP method for rotation-invariant. The LBP operator is powerful for texture description. It labels the image pixels by thresholding the surrounding pixels with comparing the value of center pixel and summing the thresholded values weighted by powers of two. The LBPlabel can be obtained with the formula in Eq. (3).

$$LBP_{P,R}(x, y) = \sum_{p=0}^{P-1} s\left(f(x, y) - f(x_p, y_p)\right) 2^p \quad (3)$$

where f(x, y) is the center pixel (red dot) of image *f* shown as Figure 10. *P* is the number of surrounding points, R is sampling radius, and s(z) is the thresholding function shown as Eq.(4).

$$s(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases} \quad (4)$$

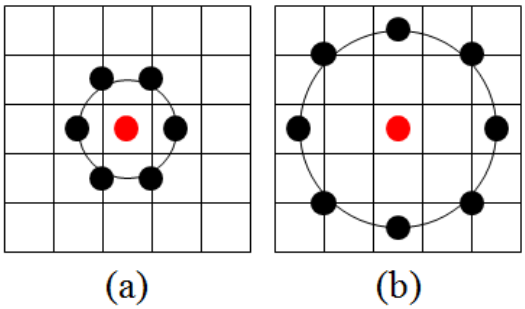


Figure 10. LBP sampling radius. (a) (P, R) = (6, 1). (b) (P, R) = (8, 2).

Furthermore, an extended LBP operator called uniform LBP [11] is proposed to describe the region texture distribution more precisely. The uniform LBP operator is constructed by considering if the binary pattern contains at most two bitwise transitions from 0 to 1 or 1 to 0 when the bit pattern is considered circular[6]. For computing the uniform LBP histogram, each uniform pattern is assigned to a specified bin and all non-uniform patterns are assigned into a single bin.The58 possible uniform patterns (all zeros, all ones, non-uniform)of (P, R) = (8, R) are shown in Figure11.

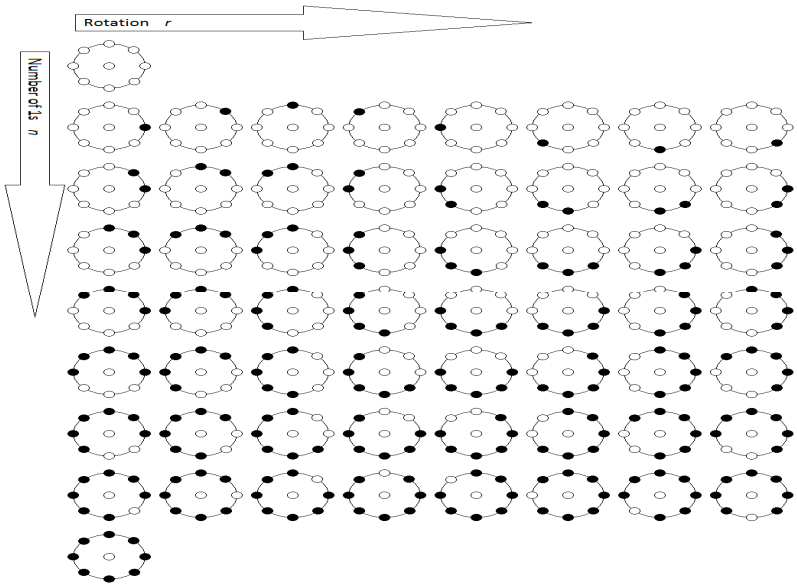


Figure 11. 58 possible uniform patternsof (P, R) = (8, R).

The uniform LBP owns a rotation invariant property. The rotation of uniform LBP is just as a horizontal shift in Figure 11 and shown in Figure 12. Based on this property, the LBP-HF [6] image feature is proposed. The LBP-HF image feature is generated by performing Fourier transform to every row in the uniform LBP histogram(except the first and the last row) to Discrete Fourier Transform to construct these features, and let $H(n, \cdot)$be the DFT of n-th row of the histogram $h_I(UP(n, r))$,, which is shown as Eq. (5).

$$H(n,u) = \sum_{r=0}^{P-1} h_I\big(U_P(n,r)\big)e^{-i2\pi ur/P} \quad (5)$$

In Eq. (5), $H(n, u)$ is the Fourier transformed histogram, n is the number of "*1*", u is the frequency, h_I is the uniform LBP histogram of the image I, $U_P(n, r)$ is the uniform LBP operator, and r denotes the row index. We apply the feature vectors consisting of three LBP histogram values (all zeros, all ones, non-uniform) and Fourier magnitude spectrum values of LBP-HFinEq. (6) to describe the texture distribution of the food ingredient image.

$$fv_{\mathrm{LBP-HF}} = [\ \ |\mathrm{H}(1,0)|, \cdots, \left|\mathrm{H}\left(1, \frac{\mathrm{P}}{2}\right)\right|, \ldots \\ \cdots, |\mathrm{H}(\mathrm{P}-1,0)|, \cdots, \left|\mathrm{H}\left(\mathrm{P}-1, \frac{\mathrm{P}}{2}\right)\right|, \\ \mathrm{h}\left(\mathrm{U_P}(0,0)\right), \mathrm{h}\left(\mathrm{U_P}(\mathrm{P},0)\right), \mathrm{h}\left(\mathrm{U_P}(\mathrm{P}+1,0)\right)\ \ ] \tag{6}$$

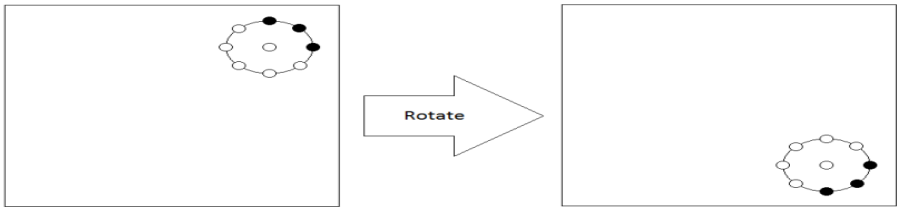


Figure12. Rotation doesn't change the row it belongs to in Figure11

2.2.3.Data Fusion and Matching

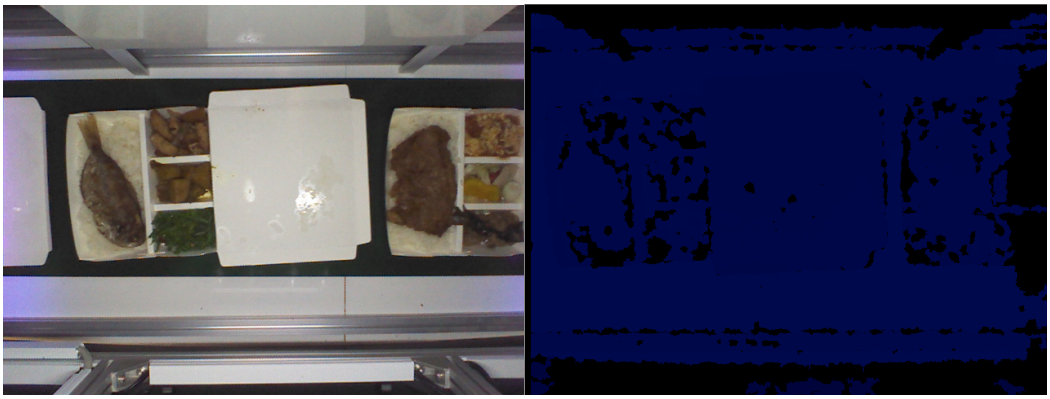
In this study, we utilize Bhattacharyya distance[12]to measure the similarity between the trained and test patterns that are described with the LBP-HF texture description and polar color histogram. Bhattacharyya distance measurement shown in Eq. (7) can be used to measure the similarity between two discrete probability distributions.

$$\mathrm{d_B(y)} = \frac{1}{\mathrm{S}}\sum_{\mathrm{b}=1}^{\mathrm{S}}\sqrt{1-\rho_{\mathrm{B}}[\mathrm{H_b}, \mathrm{P_b(y)}]}, \tag{7}$$

where, $\rho_{\mathrm{B}}[\mathrm{H_b}, \mathrm{P_b(y)}] = \sum_{\mathrm{i}=1}^{\mathrm{m}}\sqrt{\frac{\mathrm{h_i}\cdot\mathrm{p_i(y)}}{\sum_{\mathrm{i}=1}^{\mathrm{m}}\mathrm{h_i}\cdot\sum_{\mathrm{i}=1}^{\mathrm{m}}\mathrm{p_i(y)}}}$.

2.3.Food Quantity Measurement

For the inspection of amount of food ingredient, we use depth information obtained from the depth sensor to evaluate the amount of each food ingredient. Figure 13illustrates the captured depth information used to determine the amount of food ingredient.



(a) (b)

Figure13.The captured depth information used to determine the amount of food ingredient. (a) Meal box image. (b) Depth image for (a).

3. EXPERIMENTAL RESULTS

In this section, we apply the automatic meal box detection/locating module and automatic food ingredients identification module to construct a food quality inspection system. The operation scenario is shown in Figure 14.

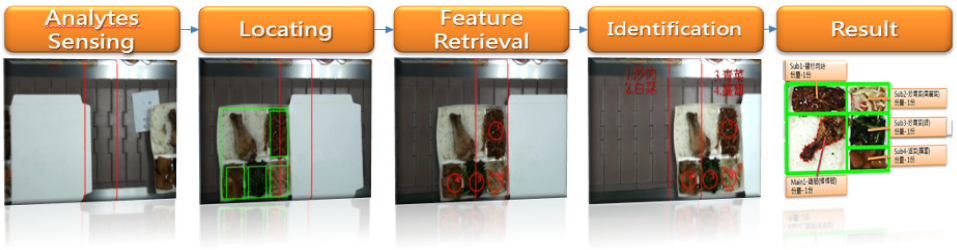
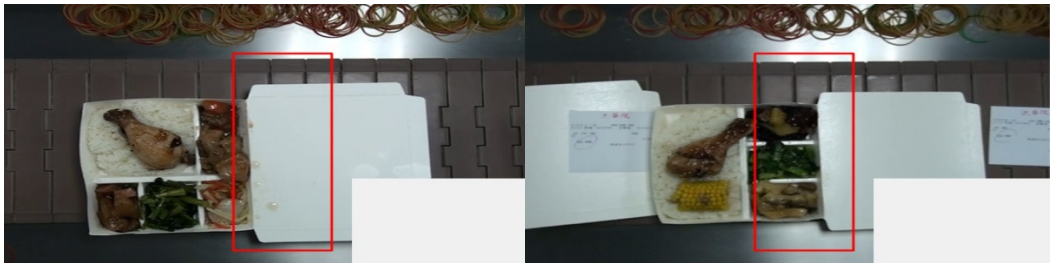


Figure 14.The system operation scenario.

The proposed food quality inspection system is implemented on the meal box dispatch conveyor of the Chinese food central kitchen. It automatically check compliance of the meal box content between customer-made meals orders of the dietician designed. This automated inspection system's operation module used Intel i5 2.2GHz CPU to analysis the contents of meal box, and it used the Microsoft Kinect camera in capture module.

The performance of food quality inspection is evaluated with two different meal boxes that are three and four food ingredients' partitions. Figure 15 shows two different meal boxes. There are 9 dishes types in the meal box including the one main dishes and 3 or 4 vice-dishes, which are shown in Figure 15.The efficiency of the meal box location detection and locating module and food ingredients identification module is listed in Table 1.



(a) (b)

Figure 15.(a) 4 vice-dish meal box. (b) 3 vice-dish meal box.

Table 1.The efficiency analysis of the meal box detection and locating.

Test Video	Meal box location detect module(1 box)	food ingredients identification module(1 box)	AVG. time for one meal box
Video 1 (4 vice-dish)	10.16 ms	116.44ms	126.60 ms
Video 2 (3 vice-dish)	9.8 ms	95 ms	114.8 ms.

Table 2 illustrates the accuracy analysis of the proposed food quality inspection system. The accuracy of meal box detection and locating is higher than 85% and the accuracy of food ingredients identification can approach 85%.

Table 2.Accuracy analysis for the food quality inspection system.

Test Video	Meal box location detect module	food ingredients identification module
Video 1 (4 vice-dish)	85.3 %	82.1 %
Video 2 (3 vice-dish)	89.6 %	89.3 %

For the amount inspection of each food ingredient, we apply the depth information to evaluate the amount of each food ingredient. The efficiencies for the meal box location detection module, food ingredients identification, and quantity estimated module are shown in Table 3.The complete average processing time of each meal box is about 1.4 second. In Table 4, the accuracy analysis is listed.

Table 3.Detection efficiency of the automated optical inspection system

Test Video	Meal box location detect module(1 box)	food ingredients identification module(1 box)	food quantity estimated module(1 box)	AVG. time for one meal box
Video 1	21.36 ms	94.1 ms	25.2ms	140.66 ms

Table 4.Detection accuracy of the automated optical inspection system

Test Video	Meal box location detect module	food ingredients identification module	food quantity estimated module
Video 1	85.3 %	82.1 %	74.2%

4. CONCLUSIONS

In the proposed system, firstly, the meal box can be located automatically with the vision-based method and then all the food ingredients can be identified by using the colour and LBP-HF texture features. Secondly, the quantity for each of food ingredient is estimated by using the image depth information. The experimental results show that the dietary inspection accuracy can approach 80%, dietary inspection efficiency can reach 1200ms, and the food quantity accuracy is about 90%. The proposed system is expected to increase the capacity of meal supply over 50% and be helpful to the dietician in the hospital for saving the time in the diet identification process.

REFERENCES

[1] Chetima, M.M. &Payeur, p. (2008) "Feature selection for a real-time vision-based food inspection system", Proc. of the IEEE Intl Workshop on Robotic and Sensors Environments, pp 120-125.

[2] Blasco, J. Aleixos, N. Molto, E. (2007) "Computer vision detection of peel defects in citrus by means of a region oriented segmentation algorithm", Journal of Food Engineering, Vol. 81, No. 3, pp535-543.

[3] Brosnan, T. & Sun, D.-W.(2004) "Improving quality inspection of food products by computer vision – a review", Journal of Food Engineering, Vol. 61, pp. 3-16.

[4] Matsuda, Y. & Hoashi, H. (2012) "Recognition of multiple-food images by detecting candidate regions." In Proc. of IEEE International Conference on Multimedia and Expo, pp 1554–1564.

[5] Yang, S. Chen, M. Pomerleau, D. Sukhankar. (2010)"Food recognition using statistics of pairwise local features."International Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, pp. 2249–2256

[6] Zhao, G. Ahonen, T. Matas, J. Pietikainen, M. (2012) "Rotation-invariant image and video description with local binary pattern features," IEEE Trans. Image Processing, Vol. 21, No. 4, pp. 1465 –1477.

[7] Suau, P., Rizo, R. Pujol, M. (2004) "Image Recognition Using Polar Histograms." http://www.dccia.ua.es/~pablo/papers/ukrobraz2004.pdf

[8] Chen, M.Y., Yang, Y.H., Ho, C.J., Wang, S.H., Liu, S.M., Chang, E., Yeh, C.H., Ouhyoung, M. (2012) "Automatic Chinese food identification and quantity estimation." SIGGRAPH Asia 2012 Technical Briefs

[9] Chang, P., Krumm, J.(1999) "Object Recognition with Color Coocurrence Histograms", IEEE International Conference on Computer Vision and Pattern Recognition,

[10] Ren, J., Jiang, X., Yuan, J., Wang, G., (2014)"Optimizing LBP Structure For Visual Recognition Using Binary Quadratic Programming", Signal Processing Letters, IEEE, On page(s): 1346 - 1350 Volume: 21, Issue: 11.

[11] Ojala, T., Pietik¨ainen, M., M¨aenp¨a¨a, T.(2002)"Multi resolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987

[12] F. Aherne, N. Thacker and P. Rockett,(1998)"The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data," Kybernetika, vol. 34, no. 4, pp. 363-368.

[13] M.R. Chandraratne, D. Kulasiri, and S. Samarasinghe,(2007) "Classification of Lamb Carcass Using Machine Vision: Comparison of Statistical and Neural Network Analyses", Journal of Food Engineering, vol. 82, no. 1, pp. 26-34.

[14] Aleixos, N., Blasco, J., &Moltó, E. (1999). "Design of a vision system for real-time inspection of oranges."In VIII national symposium on pattern recognition and image analysis. Bilbao, Spain.pp. 387–394

INTENTIONAL BLANK

USE OF EIGENVALUES AND EIGENVECTORS TO ANALYZE BIPARTIVITY OF NETWORK GRAPHS

Natarajan Meghanathan

Jackson State University, 1400 Lynch St, Jackson, MS, USA
natarajan.meghanathan@jsums.edu

ABSTRACT

This paper presents the applications of Eigenvalues and Eigenvectors (as part of spectral decomposition) to analyze the bipartivity index of graphs as well as to predict the set of vertices that will constitute the two partitions of graphs that are truly bipartite and those that are close to being bipartite. Though the largest eigenvalue and the corresponding eigenvector (called the principal eigenvalue and principal eigenvector) are typically used in the spectral analysis of network graphs, we show that the smallest eigenvalue and the smallest eigenvector (called the bipartite eigenvalue and the bipartite eigenvector) could be used to predict the bipartite partitions of network graphs. For each of the predictions, we hypothesize an expected partition for the input graph and compare that with the predicted partitions. We also analyze the impact of the number of frustrated edges (edges connecting the vertices within a partition) and their location across the two partitions on the bipartivity index. We observe that for a given number of frustrated edges, if the frustrated edges are located in the larger of the two partitions of the bipartite graph (rather than the smaller of the two partitions or equally distributed across the two partitions), the bipartivity index is likely to be relatively larger.

KEYWORDS

Eigenvalue, Eigenvector, Network Graph, Bipartivity Index, Partitions

1. INTRODUCTION

Network analysis is the study of complex relational data that capture the relationships among the members of the system. The typical goals of network analysis are to characterize the structure of the system, identify patterns of relationships, rank the constituent members based on the connections that they are part of as well as to detect communities of the members of the system. Network analysis has applications in many disciplines such as Social networks, Biological networks, Citation networks, Co-author networks, World Wide Web, Internet, Particle Physics, Electrical networks, and etc. Accordingly, the members of the system could be anything - from individual users, user groups and organizations, molecular complexes (like proteins), scholarly publications, computers and routers, websites, electrical grids and etc. The power of network analysis is to abstract the complex relationships among the constituent members simply in the form of a graph with vertices (a.k.a. nodes) and edges (a.k.a. links), which could be directed or undirected or a combination of both as well as be weighted or unit-weight edges, depending on the nature of the interactions among the members.

We model any given complex network as a graph of vertices and edges: a vertex represents an individual component of the system being modeled (e.g., users, computers, actors, protein complexes, etc) and an edge captures the interactions between them. The adjacency matrix A(G)

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 221–230, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41218

of the network graph G essentially captures the presence of edges between any two vertices. For any two vertices *i* and *j* in graph G, the entry in the i^{th} row and j^{th} column of A(G) = 1 is 1 if there is an edge from vertex *i* to vertex *j* and 0 otherwise. Depending on the nature of the interactions, the edges could be undirected (symmetric adjacency matrix) or directed (non-symmetric adjacency matrix). We encounter undirected edges when we model networks where two-way interaction is the de facto standard for any association between two nodes in a network (for example: Protein-protein interaction networks, Power grid, Science collaboration networks, Internet, Actor network, etc). Directed edges are encountered when the interactions need not be in both directions of the link between any two nodes (for example: Phone call network, Email network, World wide web, etc). Sometimes, there could be both directed and undirected edges in a network graph (e.g., metabolic networks where certain chemical reactions are reversible while certain reactions proceed in only one direction).

In this paper, we present spectral decomposition (eigenvalues and eigenvectors of the adjacency matrix of the graph) based analysis of network graphs to detect whether they are bipartite or close-to being bipartite or not and if found to be either of the two cases, we show how to predict the two partitions of the "true" or "close-to" bipartite graph. A bipartite graph is a graph wherein the set of vertices in the graph could be partitioned to two disjoint partitions and the edges of the graph connect the vertices across the partitions. In a "true" bipartite graph, there are no edges connecting the vertices within a partition. In a "close-to" bipartite graph, there may be one or few edges (called the frustrated edges) connecting the vertices within a partition.

Spectral decomposition consists of generating a continuous multi-dimensional representation (a set of eigenvalues and the corresponding eigenvectors) of the adjacency matrix of the network graph. An *eigenvector* is referred to as the vector of coordinates of the points along each axis of the multi-dimensional space and the corresponding *eigenvalue* is the length of the projection on the particular dimension. Depending on the underlying network characteristic that is to be studied, we identify the set of axes (eigenvectors) that essentially capture the variability in the data (in this paper, we make use of the smallest eigenvalue and its corresponding eigenvector): the first axis corresponds to the direction of greatest variability in the data; the second axis (orthogonal to the first axis) captures the direction of the greatest remaining variability and etc. Though the number of dimensions in the spectrum is the number of vertices in the graph, most of the variations could be captured in the first few dimensions of the coordinate system represented by the eigenvalues and the eigenvectors.

Related Work: Most of the work in the literature has focused on developing algorithms that minimize the number of edges (i.e., the frustrated edges) that need to be deleted from a graph to extract a bipartite spanning graph. Though this is an NP-hard problem for general graphs [1], for fullerene graphs (cubic 3-connected planar graph with exactly 12 pentagonal faces and an optional number of hexagonal faces), it has been found that there exists a polynomial-time algorithm [2] to determine the minimum set of edges that could be removed from fullerene graphs to extract a bipartite spanning graph. In [3], the authors developed a mathematical programming model and a genetic algorithm to determine the minimum number of frustrated edges to be removed from fullerene to extract a bipartite subgraph. In [4], the authors compute the bipartite edge frustration of a polybuckyball (a fullerene polymer) by extending the splice and link operations on the two partitions of the graph. In [5], the authors derive theoretical bounds on the maximum frustration index of a complete graph with a set of *l* and *r* vertices constituting the two partitions. Thus, most of the work in the literature is focused on minimizing the number of frustrated edges that need to be removed from selected graphs (mostly chemical compounds) to obtain a bipartite spanning graph. Our contributions in this paper are to predict the two bipartite partitions of any given graph that is hypothesized to be "true" bipartite or "close-to" bipartite as well as to analyze the impact of the distribution of the frustrated edges on the bipartivity index.

Roadmap of the Paper: The rest of the paper is organized as follows: Section 2 explains the eigenvalues and eigenvectors in greater detail and illustrates their computation with a simple example. Section 3 presents the calculation of bipartivity index on undirected graphs. Section 4 illustrates the prediction of the partitions for undirected "true" and "close-to" bipartite graphs. Section 5 illustrates the prediction of the partitions for directed "true" and "close-to" bipartite graphs. Section 6 concludes the paper.

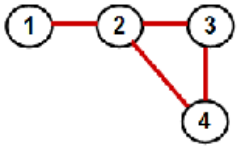
2. EIGENVALUES AND EIGENVECTORS

Spectral decomposition is a standard method to handle multivariate data in statistics and identify the directions of maximum variability [6]. The directions are called the eigenvectors and the relative importance to be given for each direction is represented by the eigenvalues. The spectrum is the collection of all the (eigenvalues, eigenvector) pairs of the multivariate data represented as a matrix. In this paper, we show that spectral decomposition of a unit-weight adjacency matrix (where the entries are either 0 or 1) of a network graph could be conducted to extract information on the extent of bipartivity in an underlying network as well as to predict the two partitions constituting the network graph.

We now illustrate an example to determine the calculation of eigenvalues and eigenvectors. Figure 1 illustrates the computation of the characteristic polynomial of an adjacency matrix for the network graph shown. The roots of the characteristic polynomial (i.e., roots of the equation $|A - \lambda I| = 0$) are the eigenvalues. Accordingly, we solve the characteristic polynomial $\lambda^4 - 4\lambda^2 - 2\lambda + 1 = 0$; the roots are $\lambda = \{2.17; 0.31; -1; -1.48\}$. The eigenvector X for an eigenvalue λ is the one that satisfies $(A - \lambda I)\, X = 0$ [7]. Note that X is a column vector with *n* rows where *n* is the dimension of the adjacency matrix A.

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \qquad I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \lambda I = \begin{bmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} -\lambda & 1 & 0 & 0 \\ 1 & -\lambda & 1 & 1 \\ 0 & 1 & -\lambda & 1 \\ 0 & 1 & 1 & -\lambda \end{bmatrix}$$



$$\text{Determinant } (A - \lambda I) = (-\lambda) * \begin{vmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{vmatrix} - 1 * \begin{vmatrix} 1 & 1 & 1 \\ 0 & -\lambda & 1 \\ 0 & 1 & -\lambda \end{vmatrix}$$

$$= (-\lambda) * \{ (-\lambda) * (\lambda^2 - 1) - 1*(-\lambda - 1) + 1*(1+\lambda) \} - 1* \{\lambda^2 - 1\}$$

$$= (-\lambda) * \{ -\lambda^3 + \lambda + \lambda + 1 + 1 + \lambda \} - \lambda^2 + 1$$

$$= \lambda^4 - 3\lambda^2 - 2\lambda - \lambda^2 + 1$$

$$= \lambda^4 - 4\lambda^2 - 2\lambda + 1 \longleftarrow \text{Characteristic Polynomial}$$

Figure 1. Characteristic Polynomial for the Adjacency Matrix of the Network Graph

We illustrate the computation of the eigenvector for eigenvalue 2.17 in Figure 2. Note that 2.17 is the largest eigenvalue for the adjacency matrix and it is called the principal eigenvalue and its corresponding eigenvector is called the principal eigenvector. For the calculation of the eigenvalues and eigenvectors of adjacency matrices used in this paper (including Figure 2), we use the website: http://www.arndt-bruenner.de/mathe/scripts/engl_eigenwert.htm. A screenshot of the results obtained for the adjacency matrix of Figure 1 is shown below in Figure 3.

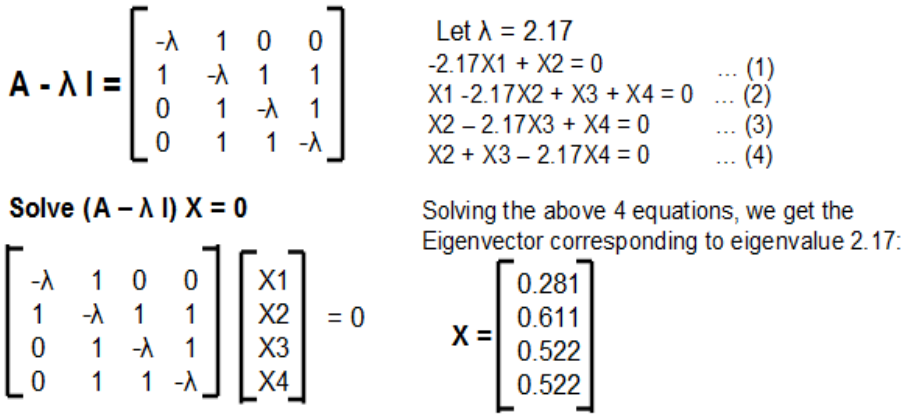


Figure 2. Calculation of the Principal Eigenvector for the Network Graph in Figure 1

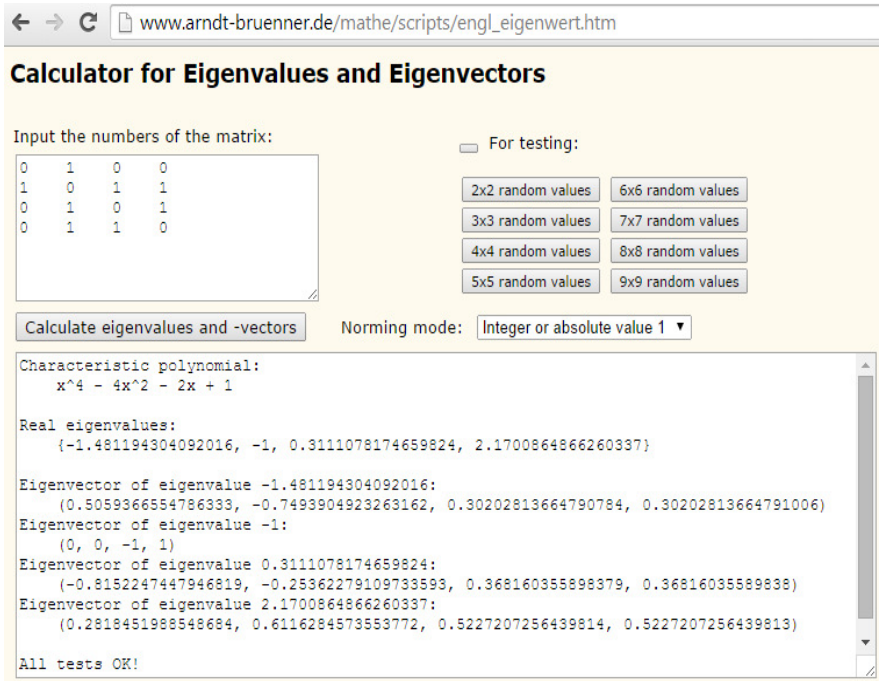


Figure 3. Online Calculator for Eigenvalues and Eigenvectors for an Adjacency Matrix

3. BIPARTIVITY INDEX

A graph G = (V, E) is said to be bipartite if the vertex set V could be divided into two disjoint sets V1 and V2 such that there are no edges connecting vertices within the two subsets and every edge

in E only connects a vertex in V1 to a vertex in V2 or vice-versa (if the graph is directed) [8]. More formally, G = (V, E) is said to be bi-partite if the two partitions V1 and V2 of the vertex set V and the edge set E are related as follows:

(i) V1 ⋎ V2 = V and V1 ⋏ V2 = Φ (empty set)
(ii) $\forall (i, j) \in$ E, either $i \in$ V1 and $j \in$ V2 or $i \in$ V2 and $j \in$ V1

Figure 4.1 illustrates a bipartite graph that has no edges within its two vertex set partitions. In reality, it may not be possible to find network graphs that are truly bipartite. There may be few edges between the vertices within the same partition. Such edges are called frustrated edges. Figure 4.2 illustrates a graph that is close to being bipartite, with the majority of the edges connecting the vertices across the two partitions but there are two frustrated edges. The eigenvalues of the adjacency matrix can be used to determine the extent of bipartitivity of a network graph G in the form of a metric called the bipartitivity index, $b_S(G)$, calculated as follows. Let $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ be the eigenvalues of the adjacency matrix of G.

$$\text{Bipartivity Index of graph G}, b_S(G) = \frac{\sum_{j=1}^{n} \cosh(\lambda_j)}{\sum_{j=1}^{n} \cosh(\lambda_j) + \sum_{j=1}^{n} \sinh(\lambda_j)}$$

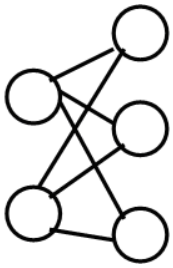


Figure 4.1. A "True" Bipartite Graph

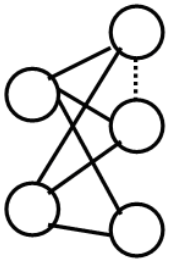
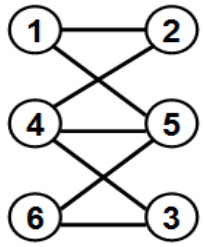


Figure 4.2. A "Close-to" Bipartite Graph

Figure 4. Examples of "True" and "Close-to" Bipartite Graphs

The calculation of the bipartivity index for a "true" bipartite graph and for a "close-to" bipartite graph are shown in Figures 5 and 6 respectively. We can notice that for a "true" bipartite graph, the $\sinh(\lambda_j)$ values in the formula for the bipartivity index add to 0, resulting in the bipartivity index being 1 for such graphs. On the other hand, for a non-bipartite graph, the sum of the $\sinh(\lambda_j)$ values contribute a positive value - leading to an increase in the value of the denominator compared to the numerator in the formula for bipartivity index. Thus, the bipartivity index for a non-bipartite graph is always less than 1; if the $b_S(G)$ values of graph G is closer to 1, we call such graphs as "close-to" bipartite, as the one in Figure 6 (where edge 2 - 4 is removed from the graph in Figure 5 and edge 1 - 4 is added as a frustrated edge).



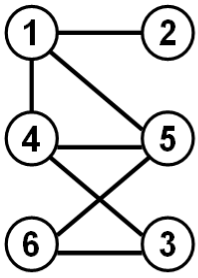
Eigenvalue, λ	cosh(λ)	sinh(λ)
-2.414	5.634	-5.545
-1	1.543	-1.175
-0.414	1.087	-0.426
0.414	1.087	0.426
1	1.543	1.175
2.414	5.634	5.545
Total	16.528	0

$$b_S(G) = \frac{\sum_{j=1}^{n} \cosh(\lambda_j)}{\sum_{j=1}^{n} \cosh(\lambda_j) + \sum_{j=1}^{n} \sinh(\lambda_j)}$$

$$b_S(G) = \frac{16.528}{(16.528 + 0)} = 1.0$$

Figure 5. Bipartivity Index Calculations for a "truly" Bipartite Graph

Figure 7 illustrates the impact of the number of frustrated edges and their location on the bipartivity index values for several sample network graphs. The bipartivity index of the graphs decreases with increase in the number of frustrated edges that connect vertices within the same partition. We can observe that for a given number of frustrated edges, a larger value of the bipartivity index is observed for graphs that have a relatively larger number of frustrated edges in the larger partition vis-á-vis the smaller partition.



Eigenvalue, λ	cosh(λ)	sinh(λ)
-2	3.762	-3.627
-1.562	2.489	-2.279
0	1	0
0	1	0
1	1.543	1.175
2.562	6.519	6.442
Total	16.313	1.711

$$b_S(G) = \frac{\sum_{j=1}^{n} \cosh(\lambda_j)}{\sum_{j=1}^{n} \cosh(\lambda_j) + \sum_{j=1}^{n} \sinh(\lambda_j)}$$

$$b_S(G) = \frac{16.313}{(16.313 + 1.711)} = 0.905$$

Figure 6. Bipartivity Index Calculations for a "close-to" Bipartite Graph

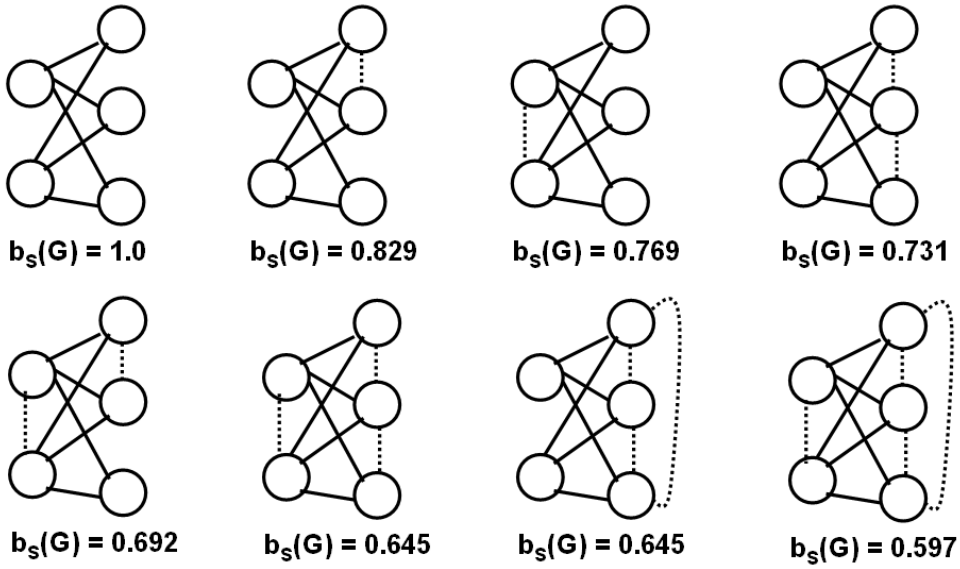


Figure 7. Impact of the Number of Frustrated Edges and their Location on the Bipartivity Index of Network Graphs

4. PREDICTIONS OF THE PARTITIONS IN A UNDIRECTED BIPARTITE GRAPH

We now illustrate how to predict the two partitions of a "true" or "close-to" bipartite graph. For this purpose, we will make use of the smallest of the eigenvalues and its corresponding eigenvector, hereafter referred to as the bipartite eigenvalue and the bipartite eigenvector respectively. The bipartite eigenvalue is most likely a negative value in "true" or "close-to" bipartite graphs. The bipartite eigenvector is likely to comprise of both positive and negative entries. The node IDs whose entries in the bipartite eigenvector are of the positive sign constitute one of the two partitions and those of the negative sign constitute the other partition. The above approach has been found to accurately predict the two partitions of a "true" bipartite graph, as shown in Figure 8. However, for "close-to" bipartite graphs, the partitions predicted (using the smallest eigenvalue and its corresponding eigenvector) may not be the same as the partitions expected (hypothetical partitions) of the input graph whose adjacency matrix had been used to determine the eigenvalue and the eigenvector. Nevertheless, the predicted partitions of the "close-to" bipartite graphs and the hypothetical partitions of the original input graph contribute to the same bipartivity index value. This shows that two "close-to" bipartite graphs that physically look similar (i.e., same set of vertices and edges connecting the vertices), but are logically different (i.e., differ in the partitions) would still have the same bipartivity index; the difference gets compensated in the number of vertices that form the two partitions and/or the distribution of the frustrated edges across the two partitions. Note that the predictions of the partitions get less accurate as the bipartivitiy index gets far lower than 1.

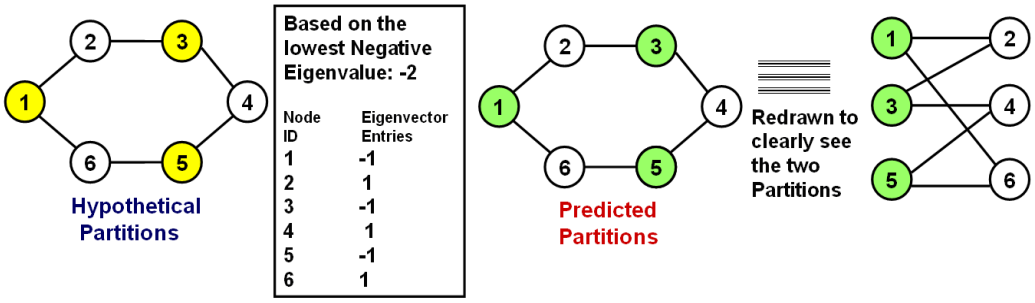


Figure 8. "True" Bipartite Graph: Predicted Partitions Match with the Hypothetical Partitions of the Input Graph

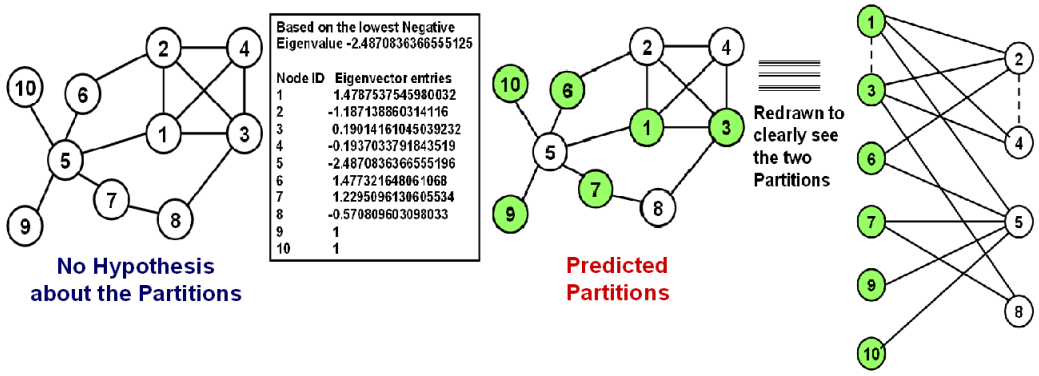


Figure 9. "Close-to" Bipartite Graph: Predicted Partitions Appear to be the Same as that of the Input Graph

Figure 9 illustrates the prediction of the partitions of a "close-to" bipartite graph that looks complex enough for one to initially hypothesize the two partitions; hence, we assume the predicted partitions are the same as what is as expected of the original input graph. However, Figure 10 illustrates an example where the predicted partitions of a "close-to" bipartite graph are of unequal sizes with two frustrated edges (whereas the input graph is hypothesized to have two equal-sized partitions with one frustrated edge, as shown); but, both the graphs have the same bipartitivity index. The predicted "close-to" bipartite graph consists of a larger partition with four vertices and a smaller partition with two vertices; there are two frustrated edges in the larger partition and none in the smaller partition. The input "close-to" bipartite graph for this illustration has three vertices in each of the two partitions, with a frustrated edge in one of the two partitions. This example reiterates our earlier assertion that for two "close-to" bipartite graphs that physically look the same and have the same bipartivity index, there could be logically different sets of partitions: a topology with equal number of vertices in both the partitions and fewer frustrated edges could offset for a topology with a larger partition containing relatively larger number of frustrated edges.

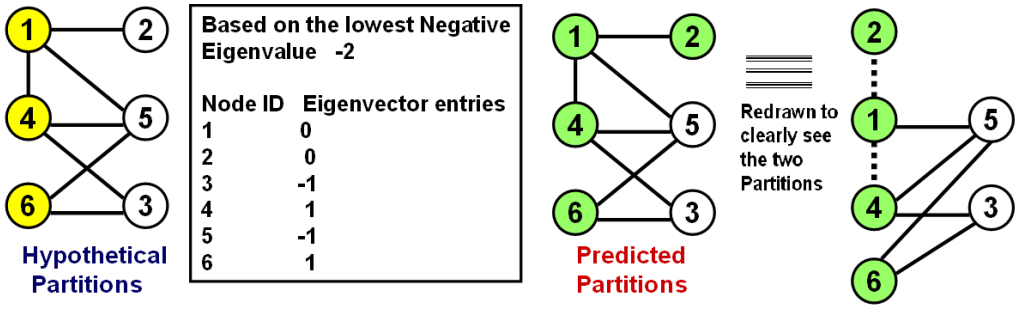


Figure 10. "Close-to" Bipartite Graph: Predicted Partitions do not Match with the Hypothetical Partitions of the Input Graph

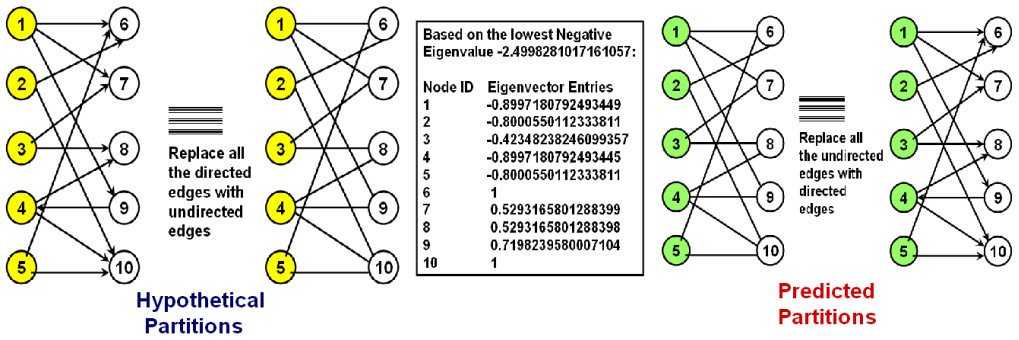


Figure 11. Predicting the Partitions of a "True" Bipartite *Directed* Graph: Predicted Partitions Match with the Hypothetical Partitions of the Input Graph

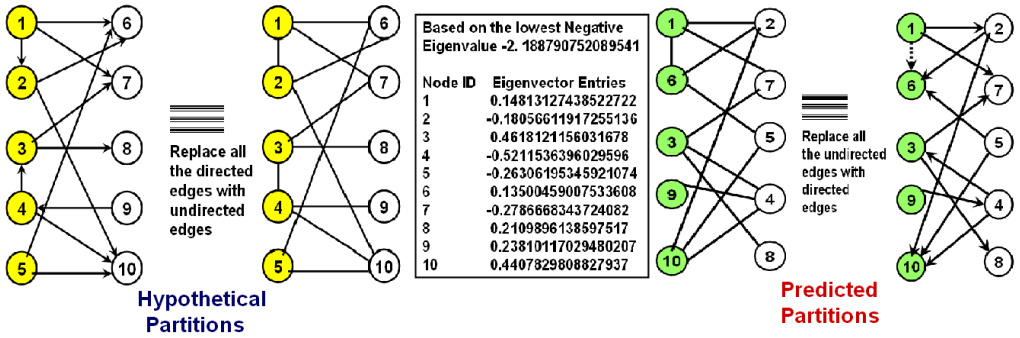


Figure 12. Predicting the Partitions of a "Close-to" Bipartite *Directed* Graph: Predicted Partitions do not Match with the Hypothetical Partitions of the Input Graph

5. PREDICTION OF THE PARTITIONS IN A DIRECTED BIPARTITE GRAPH

To predict the two partitions of a directed bipartite graph, we can still use the same approach as explained above. We need to transform the directed graph to an undirected graph (replace all the directed edges with undirected edges), determine the bipartite eigenvalue and bipartite eigenvector of the undirected graph, and predict the constituent vertices of the two partitions based on the sign of the entries (corresponding to these vertices) in the bipartite eigenvector. Finally, we restore the directions of the edges. If the input directed graph is "truly" bipartite, the predicted partitions will be the same as that hypothesized for the input graph (refer the example shown in Figure 11). However, for "close-to" bipartite directed graphs, the predicted partitions need not be the same as that of the hypothetical partitions of the input graph; but, as long as the set of vertices and edges (including the directions of the edges) remain unaltered, the bipartivity index of the "close-to" bipartite graph will remain the same with both the hypothetical expected partitions and the predicted partitions. Figure 12 illustrates an example wherein the set of vertices constituting the predicted partitions is observed to be different from the hypothetical partitions expected of the input graph. The hypothetical partitions contributed to two frustrated edges, whereas the predicted partitions contributed to only one frustrated edge. Nevertheless, since the set of vertices and the set of edges are the same for both the input graph and the graph based on the predicted partitions, the bipartivity index value remains the same.

6. CONCLUSIONS

The paper demonstrates the application of the eigenvalues and eigenvectors to analyze the bipartivity of both undirected and directed graphs. We observe that for a given number of frustrated edges, the bipartivity index is more likely to be larger if more of these edges are located in the larger of the two partitions of the bipartite graph. For "close-to" bipartite graphs, we observe the predicted partitions of the vertices to be different from that of the hypothetical partitions of the input graph; but nevertheless, since the set of vertices and set of edges constituting the bipartite graphs do not change, the bipartivity index remains the same for both the input and predicted graphs. In other words, for a given number of vertices and edges, there could be more than one instance of a bipartite graph (i.e., there could exist one or more combinations of the two partitions) that could have the same bipartivity index value. The above argument holds good for both directed and undirected bipartite graphs.

REFERENCES

[1] Z. Yarahmadi, T. Doslic and A. R. Ashrafi, "The Bipartite Edge Frustration of Composite Graphs," Discrete Applied Mathematics, vol. 158, no. 14, pp. 1551-1558, July 2010.
[2] T. Doslic and D. Vukicevic, "Computing the Bipartite Edge Frustration of Fullerene Graphs," Discrete Applied Mathematics, vol. 155, no. 10, pp. 1294-1301, May 2007.
[3] Z. Seif and M. B. Ahmadi, "Computing Frustration Index using Genetic Algorithm," Communications in Mathematical and in Computer Chemistry, vol. 71, pp. 437-443, 2014.
[4] Z. Yarahmadi, "The Bipartite Edge Frustration of Extensionof Splice and Link Graphs," Applied Mathematics Letters, vol. 23, pp. 1077-1081, 2010.
[5] G. Bowlin, "Maximum Frustration in Bipartite Signed Graphs," The Electronic Journal of Combinatorics, vol. 19, no. 4, #P10, 2012.
[6] C. Yang, B. A. Poser, W. Deng and V. A. Stenger, "Spectral Decomposition of Susceptibility Artifacts for Spectral-Spatial Radiofrequency Pulse Design," Magnetic Resonance in Medicine, vol. 68, no. 6, pp. 1905-1910, December 2012.
[7] D. C. Lay, Linear Algebra and its Applications, Pearson, 4th edition, May 2011.

TARGET-ORIENTED GENERIC FINGERPRINT-BASED MOLECULAR REPRESENTATION

Petr Skoda and David Hoksza

Faculty of Mathematics and Physics,
Charles University in Prague, Prague, Czech Republic
skoda@ksi.mff.cuni.cz
hoksza@ksi.mff.cuni.cz

ABSTRACT

The screening of chemical libraries is an important step in the drug discovery process. The existing chemical libraries contain up to millions of compounds. As the screening at such scale is expensive, the virtual screening is often utilized. There exist several variants of virtual screening and ligand-based virtual screening is one of them. It utilizes the similarity of screened chemical compounds to known compounds. Besides the employed similarity measure, another aspect greatly influencing the performance of ligand-based virtual screening is the chosen chemical compound representation. In this paper, we introduce a fragment-based representation of chemical compounds. Our representation utilizes fragments to represent a compound where each fragment is represented by its physico-chemical descriptors. The representation is highly parametrizable, especially in the area of physico-chemical descriptors selection and application. In order to test the performance of our method, we utilized an existing framework for virtual screening benchmarking. The results show that our method is comparable to the best existing approaches and on some data sets it outperforms them.

KEYWORDS

Virtual screening, Molecular representation, Molecular fingerprints

1. INTRODUCTION

The main method to identify new leads in the drug discovery process has traditionally been medium or high-throughput screening (HTS). In this experimental process, a large number of chemical compounds can be screened against a specific target to identify compounds which trigger a response in this target. Some of the HTS approaches can guarantee throughput up to about 100.000 compounds per second [1] by using the combinatorial libraries. Obviously, the throughput in such cases is not an issue anymore. However, management of such large libraries can be difficult and economically unfeasible since every new compound brought into the screening process increases its price.

The in-silico answer to the growing size of chemical databases is the so-called high-throughput virtual screening (HTVS). It allows fast screening of large libraries, which may contain up to tens of millions chemical compounds, without the need of physically own the compounds. An additional bonus which relates to the HTVS is the ability to screen even virtual libraries. I.e., one

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 231–244, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41219

can easily predict bioactivity of compounds residing in not yet well explored parts of the chemical space [2].

While the limits of HTS are given by the technology, the HTVS, since it is a simulation of a real world approach, is limited by the available information about ligands [3]. The available knowledge then dictates how HTVS is utilized. Since unlike HTS, HTVS can suffer by both false positives and false negatives it is commonly used as a pre-step to the standard HTS in the early-stages in the drug discovery pipeline. The HTVS is used to prioritize large chemical libraries which narrows down the set of compounds to be forwarded to HTS. Usefulness of complementing HTS with HTVS has been supported by several studies [4], [5]. The virtual screening approaches can be classified as ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) [6], [7]. The choice of which approach to utilize depends on information about the task at hand. If we know the three-dimensional structure of the biological target we can use SBVS methods [8], [9]. The SBVS is based on docking and includes two steps: positioning the ligand into the target active site (docking) and scoring the pose. However, this information is often not available in sufficient quality or it is not available at all. In such a case the ligand-based virtual screening method is the method of choice.

1.1 Ligand-based virtual screening

In LBVS, only the information about known bioactive ligands (triggering response in the given biological target) is required. The LBVS is built around the concept that similar structures carry out similar functions more often than dissimilar ones. This assumption is based on the shape and physicochemical complementary of the ligand and target commonly called key-and-lock principle [10] or similar property principle [11]. Thus, given the known active (and possibly also inactive) compounds LBVS methods prioritize compounds that are more likely to have desired functionality/features, based on the similarity to the known active molecules.

In the first step of LBVS, a computer-based representation is calculated for the known bioactive ligands as well as for all the molecules in a library to be searched for new bioactive compounds. In the second step, the representation of ligands can be aggregated and used as a query or individual representations are used directly for searching the library. As the last step, the library is sorted with respect to the similarity to the query ligand(s). It is assumed that the high-scoring compounds bind to the target with high probability due to the similarity principle.

One can come up with various classification of LBVS approaches. For example, Taboureau at. al. [7] divide LBVS into five classes based on the utilized molecular features: alignment-based, descriptor-based, graph-based, shape-based and pharmacophore-based.

While the methods might differ in the specifics of how to approach the identification of bioactive compounds, most of them employ a feature extraction step where the molecular descriptors are identified and encoded into some kind of representation. This is then used as a representation of the molecule in the virtual screening. Among the commonly used features are those which reflect structure or capture computed or experimentally measured physico-chemical properties. Currently, there exists a plethora of descriptors to be utilized in virtual screening [12], [13]. They differ not only in their semantics but also in the computational complexity. The excess of descriptors is the consequence of the fact that none of the descriptors can be generally declared as superior to the rest. The features discriminating active and inactive compounds simply depend on the specific target which varies in every screening campaign. It follows that it is vital to the success of a virtual screening campaign to capture such features which represent the molecules well in terms of their discriminative capability. This is the main motivation for our work. It is out of question that the correct choice of features greatly influences the outcome of a virtual

screening campaign. However, we moreover believe that the choice of descriptors should be context-aware that is it should be dependent on the investigated target. Therefore, in this paper we propose a general framework which allows the user to parameterize the molecular representations.

1.2 Fingerprints

A common type of descriptors are the 2D fingerprints (fingerprints) capturing the structure of given chemical compound in the form of a bitstring. Every structural feature is mapped to a position in the string. Such representation is suitable for large-scale virtual screening campaigns since it allows fast comparison of two molecules (bitstrings).

Thus, the main idea behind the fingerprints is to encode the existence of a given (structural, pharmacophore, …) feature to a position in a bitstring. The features to be encoded commonly include molecular fragments which are continuous substructures of a given molecule. There are two main approaches to fragment extraction: path-based (Topological Torsions fingerprints), neighbourhood-based (Circular Fingerprints or Extended connectivity fingerprints).

The Topological torsions fingerprints [14] (TT) use paths of length four (quaternions). The information about types, nonhydrogen connections and number of pi-electrons is used to calculate the index of given path.

In Extended Connectivity Fingerprints (ECFPs) and Functional Connectivity Fingerprints (FCFPs) an atom is described in terms of its neighbouring atoms up to a certain radius. Hert [15] has shown that such descriptors can be effective in similarity searching applications. The extended connectivity of an atom is calculated using a modified version of the Morgan algorithm [16] where the atom code is combined with the codes of its neighbours to establish the final atom description.

To map a fragment into a position in the bitstring representation, a mapping function needs to be utilized. The simplest solution is the dictionary-based approach where a predefined dictionary of fragments and their mapping into the bitstring is utilized. However, this allows to represent only a limited set of fragments in the bitstring. Another solution is to map every possible fragment into a constant-sized bitstring. However, since the size of a bitstring representation uses to be an order of magnitude smaller in comparison to the number of all possible fragments, typically a modulo function is applied. This allows to obtain a bitstring position for every possible fragment. On the other hand, two different fragments with different indexes can be mapped to the same position in a bitstring. This situation is called the collision. The fingerprints that utilize this approach form the family of hashed fingerprints.

2. METHOD OUTLINE

In this work we introduce vector fingerprints (VectorFp), a new approach to the representation of chemical compounds and their comparison. As mentioned above, our goal is to provide a modular molecular representation for LBVS allowing to be parametrized based on the task at hand. The basis of VectorFp molecular representation form structural fragments. But unlike other descriptors, VectorFp allows the fragments to be labeled by user-defined physico-chemical properties. Moreover, the representation was designed with the emphasis on the ability to use it with existing similarity measures for bitstrings. Thus, VectorFp is designed as a generic representation that needs proper parametrization before it can be used.

2.1. VectorFp structure

In order to maintain compatibility with existing fingerprint methods we decided to choose the bitstring as the representation for VectorFp. The advantage is that we can utilize existing well-established similarity measures, LBVS processes, and benchmarking platforms in order to get comparison of our method to the other fingerprints.

VectorFp (Figure 1) is basically an array (outer array), where each cell represents one (or more in case of a collision) fragment(s). Each cell of the main outer array contains another array (inner array). The purpose of the inner array is to store the selected descriptors of a respective fragment(s). As mentioned, these descriptors are physico-chemical properties of fragments that are converted into bitstring representation.

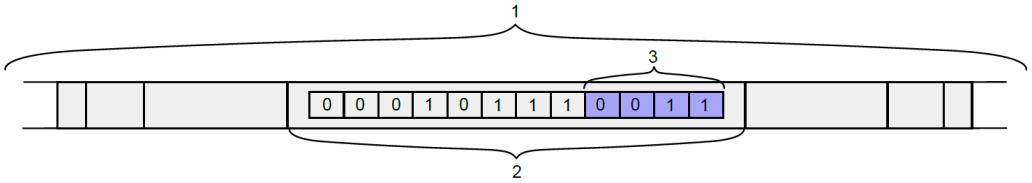


Figure 1. Structure of VectorFp. 1 - outer array, 2 - cell with inner array, 3 - bits representing single descriptor

Generally, physico-chemical descriptors can take various ranges of values being typically integer or float data types. The process of conversion of descriptors into a bitstring is secured by so called conversion methods. In our current implementation we use the same conversion method for all descriptors. It gets minimum and maximum value for a given descriptor and then uses binning which results in an integer value to be used as the descriptor value to be stored. The integer value is then encoded into a bit array using unary encoding. The binning is the formation of a set of disjoint intervals (bins) that represent the possible values. The bin index is finally encoded into a binary representation. In VectorFp we decided to use unary coding. The choice of unary coding instead of, e.g. classical binary coding, stems from the typical choice of similarity functions used when comparing bitstring molecular representations. The most commonly used similarity functions basically assess similarity to a pair of bit strings based on the number of common and differing bit positions. These measures assume that the bits are independent which holds when every bit corresponds to the existence or nonexistence of a molecular substructure. However, when the binary image of a substructure spans multiple bit positions (inner array) the positions are dependent. Using the binary coding with such similarity measure is then not valid. Let us consider a situation when the binary representation takes 4 bits. Then if the distance/similarity is based on the number of common bits bin 4 (0100) is from bin 1 (0001) in the same distance as, e.g., bin 2 (0010). Which should not hold since the bin indexes approximate quantitative characteristics. However, when using unary coding bin 1 gets the code 1000, bin 2 gets the code 1100 and bin 4 gets the code 1111. Then, using the same similarity measure, bin 2 is more similar to bin 1 than bin 4 as one would expected.

2.2. VectorFp generation

1. The VectorFp representation computation for a given molecule consists of five main steps:
2. Extract fragments from the molecule and compute indexes (positions in the bitstring representation) for those fragments.
3. For each fragment compute its physico-chemical descriptors.
4. Convert all fragments descriptors into bitstrings.
5. Create the fragment bitstring representation from its respective descriptor representations.

6. Combine the individual fragments representations (bitstrings) together and assemble the representation of the molecule. The representations of fragments are stored into cells determined by the index computed in step 1.

As the VectorFp size is limited, the index computed in step 1 must be modified by application of the modulo operation (hashing). As a consequence a collision may occur. In order to solve collisions VectorFp utilizes the bitwise logical or to merge representations of multiple fragments together. The advantage of this method is simplicity and the fact that the results (fragment bitstring) are the same for different permutations of the same fragments. The drawback of selected approach is, that created fragment representation does not have to represent existing fragment. This can be problem during similarity comparison of two VectorFps. If both molecules (their VectorFps representations) have the same fragments in single cell, then everything is in order, but if one molecule has difference number of fragments in given cell then the other molecule, for example one and two, the comparison still compare fragment representation to fragment representation. In this case we compare existing fragment to some imaginary aggregated fragment.

3. PARAMETERIZATION

From the description of VectorFp one can notice that there are places where the approach is not fully specified: fragment extraction, descriptor selection and conversion. The named areas and some more create space for parameterization of VectorFp. VectorFp can be seen as a generic representation or frame. The parameterization determines the efficiency of the final VectorFp-based molecular representation.

3.1 VectorFp size

One of the parameters is the size of VectorFp. The size is determined by two variables: size of the inner array and the number of cells in the outer array. The final size of vectorFp representation is therefore *size of inner array * size of outer array*. So for example if we use 1024 cells for the outer array, then a 4 bit increase of the inner array size will result in 4096 bit increase of the resulting representation size.

3.2 Fragment extraction

In the current implementation we utilize RDKit's [17] algorithm to extract the fragments from a molecule get their positions in the bitstring. The algorithm uses RDKit's Morgan Fingerprint which is based on the Morgan algorithm. Morgan Fingerprints use the following features to calculate a fragment's position in the bitstring: donor, acceptor, aromatic, halogen, basic, acidic. The RDKit provides the possibility to modify this feature list and thus change the fragment indexes. This can be also viewed as a possible parameterization of VectorFp. Another possibility is to use paths (like TT fingerprints) instead of neighbourhoods.

3.3 Fragment representation

Each fragment is represented by an inner array (bitstring). The size of this array determines how many information can be stored about each fragment. By setting the size of the inner array to one we get the classical fingerprints. The selection of used descriptors, conversion method and number of bits in inner array is also part of the parameterization. The descriptor selection and conversions are in our opinion the most important parts of the parameterization and have a great influence on the performance of the method. The selected descriptors include, for example, the number of heavy atoms, logP, the presence of a fragment or, in an extreme case, other fingerprint

can be used as a fragment's descriptor and inserted into VectorFp. There is also the possibility to stress certain descriptor by multiplying its value. For example, let us have two different descriptors, we use them both in our parameterization but we replicate one of them. In this case, the replicated descriptor has more weight and can be seen as the main one. The second one (nonreplicated) descriptor can serve as a fine tuning mechanism.

4. EXPERIMENTS

For experimental evaluation we used the recently published framework for benchmarking LBVS approaches by Riniker et al. [18]. The framework is written in Python [19] and uses RDKit [17] as the underlying chemical framework. It comes with a predefined set of fingerprints, similarity methods (Dice, Tanimoto, Cosine, Russel, Kulczynski, McConnaughey, Manhattan, RogotGoldberg) and quality measurement methods (Area Under Curve (AUC) of Receiver Operating Characteristic curve (ROC), Enrichment Factor (EF), Robust Initial Enhancement (RIE) [20], Boltzmann-Enhanced Discrimination of ROC (BEDROC) [21]). The framework simulates LBVS on pooled targets from three data sets representing 88 targets in total. The three data sets include Database of Useful Decoys (DUD) [22], ChEMBL [23] and Maximum Unbiased Validation (MUV) [24]. For each target a set of known actives and inactives (decoys) is available. As the framework aims to high reproducibility of experiments it also contains a predefined random selection of actives and decoys. Thanks to that, the simulation of LBVS is deterministic and can be easily reproduced by any researcher.

However, one of the drawbacks of the framework is that it is designed to use the same method with the same parameterization for all the data sets. There is no learning phase per dataset. Such phase could be useful for benchmarking of methods including a learning phase [25]. The absence of learning phase influences performance of our method in a negative way as our method needs a proper parameterization that differs based on the task (dataset) at hand. Still, we decided to not modify the benchmarking platform and to use a single parameterization over all data sets as the determination of the right parameterization is not the goal of this article. The problem of correct parameterization and feature selection is a separate topic.

Riniker et al. [18] recommend to use at least two different benchmarking methods, for example AUC and BEDROC as the AUC alone is considered to be insufficiency sensitive. On the other hand, the advantage of the AUC in comparison to some other methods is that it is non-parametric. Thus, it can be easily used to give a basic idea about the performance of tested method especially in a large scale evaluation. From this reason, we decided to show only AUC values in the following experimental evaluation.

4.1. Comparison to existing methods

In this section, we presents the comparison of *VectorFp* with other fingerprints from selected benchmarking framework. We used *VectorFP* with the best found parameterization (aggregated over all targets). However, we emphasize that the *VectorFp* performance strongly depends on the selected parameterization (see section IV-B) and since the parameterization optimization is a hard (and separate) problem, there is still room for improvement. Moreover, in this comparison we use a single parameterization for all targets which is not the optimal and intended use of *VectorFp*, but we find it useful in order to get a rough comparison with the other existing methods. To denote the other fingerprints we use abbreviations from the original article [18] containing also the details about the remaining fingerprints.

The best parameterization we obtained in our experiments in terms of average *auc* (average of *auc* over all data sets) was *nHBDon_Lipinski,nN*. This parameterization utilizes two descriptors –

nHBDon_Lipinski and *nN*, where each descriptor occupies 16 bits in the final representation. We denote this parameterization further in the text as *vectorFp*.

As already stated, the VectorFp is designed as a generic representation that should be rather used with parameterization based on the given task. In order to demonstrate the potential of *VectorFp*, we defined virtual VectorFp (*vVectorFp*). To get the results for *vVectorFp*, we select the best tested parameterization for every dataset. Thus, *vVectorFp* can be understood as *VectorFp* with an oraculum that gives us the best encountered parameterization for given target.

As for the source of descriptors for labelling the extracted fragments, we used the PaDEL [26] tool. PaDEL is capable of generating about 770 2D descriptors that can be easily utilized in *VectorFp*. To convert the descriptor values into the bitstring in the inner arrays of *VectorFp* we use the binning and unary coding.

As the results show (Table 1.), *vectorFp* (with the *nHBDon-Lipinski*,*nN* parametrization) is, in terms of *auc*, the best fingerprint for 8 out of the 88 data sets and it ends up on position 9.966 on average. The best obtained average position is 8.092 reached by the *TT* fingerprint.Thus, although the single parameterization is used for multiple data sets, it is clearly comparable with the best existing approaches. On some data sets, our method is superior to all the other methods. The performance differs throughout all the data sets (Table 2.).

Table 1. Aggregated performance statistics of *vectorFp* and *vVectorFp* with respect to other fingerprints

name	average AUC	number of best results	average position
tt	0.8034	12	8.09
hashap	0.7701	11	14.20
rdk6	0.7821	10	12.55
vectorFp	0.7890	8	9.97
laval	0.7798	7	12.91
avalon	0.7755	7	14.03
rdk7	0.7407	6	17.62
ap	0.7914	5	10.25
hashtt	0.7973	4	9.31
rdk5	0.7827	3	12.25
lfcfp6	0.7631	3	14.71
fcfp2	0.7457	3	18.17
ecfc6	0.7795	3	11.79
fcfp4	0.7643	2	14.71
fcfc6	0.7625	2	15.10
lfcfp4	0.7620	2	15.23
lecfp4	0.7606	2	15.03
lecfp6	0.7581	2	16.03
fcfp6	0.7657	1	14.59
ecfp2	0.7522	1	17.68
fcfc2	0.7435	1	19.55
maccs	0.7333	1	20.08
ecfc4	0.7798	0	11.61
ecfc2	0.7739	0	13.68
fcfc4	0.7603	0	15.77
ecfp4	0.7582	0	16.03
ecfp6	0.7573	0	16.68

ecfc0	0.7340	0	20.43
ecfp0	0.6463	0	27.91
vVectorFp	0.8174	31	4.37

As can be seen most of the tested fingerprints perform reasonably well, in comparison to the others, on at least one dataset. Out of the three data sets, the MUV dataset shows up to be the hardest for *vVectorFp* as in 3 cases it performs strongly under average. However, the MUV dataset is the most difficult for every tested fingerprint. The goal of the MUV design is to generate sets with a spatially well distributed active and decoy molecules in a simple descriptor space. Moreover, another goal is to evenly distribute actives among the decoys which makes the MUV dataset difficult for virtual screening. The best tested parameterization for MUV shows up to be *naAromAtom16*,*ETA_BetaP_s16*,*minHsNH2* with the average *auc* on MUV being 0.6258 compared to *vectorFp* having the average *auc* of 0.6214.

As the VectorFp in fact utilizes one of the extended connectivity fingerprints (*ecfp*) as the underlying fingerprint, we were interested how it compares to the performance of other fingerprints from the same family. From this perspective our method performs well and outperforms most fingerprints from this family.

Our method was in term of average *auc* over all the data sets outperformed by *tt*, *hashtt* and *ap* fingerprints. All those fingerprints are based on different fragments than used in current version of VectorFp. *tt* and *hashtt* use paths of length four while *ap* use atom pairs. This suggest that the change of fragment extraction process (underlying fingerprint) may improve the performance of *VectorFp*.

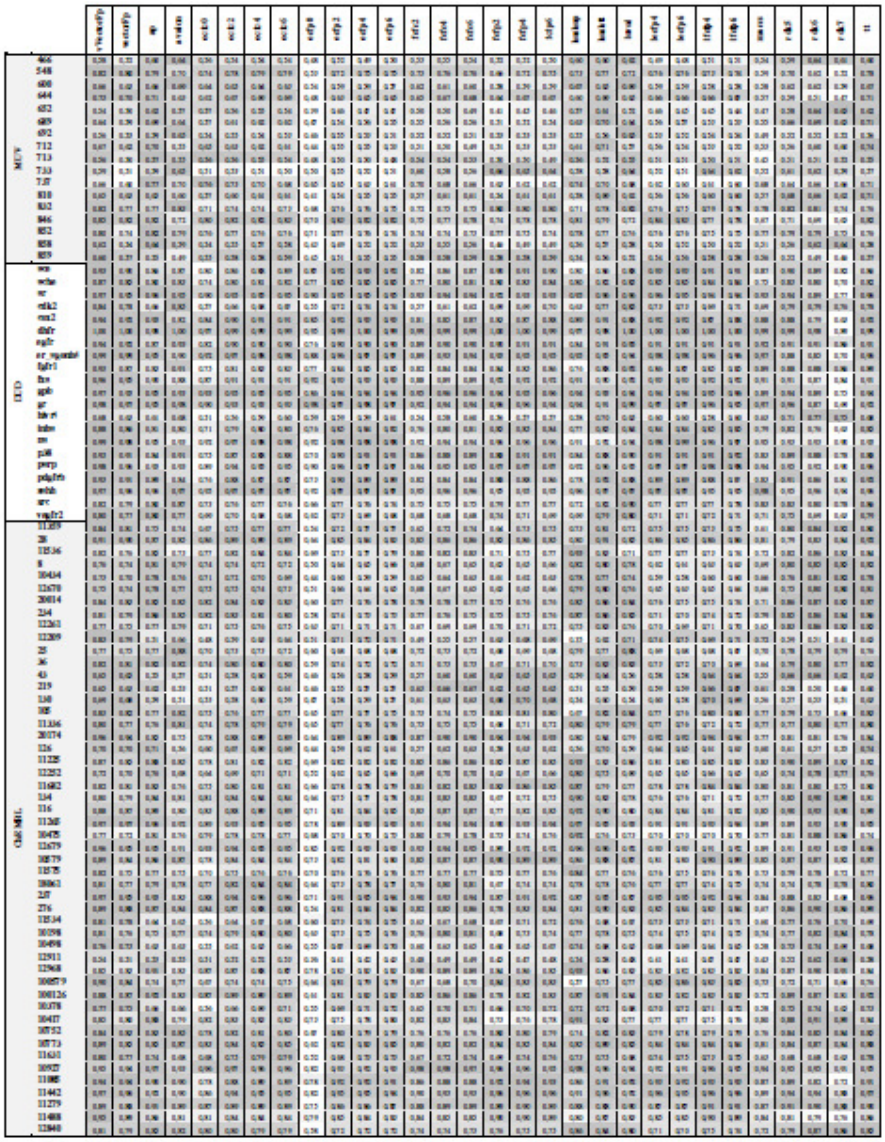
Notice, that *vVectorFp* is also included in the comparison. As it is not based on a single parameterization, the values presented in Table 1. were computed without the *vVectorFp*, and at the end the *vVectorFp* was added. Thus *vVectorFp* results did not influence the positions of other approaches. The vVectorFp outperforms all other methods in all the presented evaluation criteria (average *auc*, number of best results). We believe that this demonstrates the potential of *VectorFp* if parameterized properly. We emphasize again that there may be a better parameterization as we tested only a very limited subset of all possible parameterizations.

4.2. Parameterization

As a part of our experiments we systematically tested hundreds of different parameterizations focusing on various descriptors provided by PaDEL (see above). Although there are more ways of how to parameterize VectorFp, here we focused on descriptor selection only being the most result influencing part of the parameterization.

To test how the amount of used descriptors per fragment influences the discriminative power of the molecular representation, we started with just one descriptor per fragment and then added more. In the preparation phase, we extracted all fragments for all chemical compounds in every data sets of the benchmarking platform. For each fragment we computed all descriptors available in PaDEL. These descriptors were the subject of a basic descriptor analysis before running the experiments themselves. The goal of the analysis was to remove descriptors which clearly did not have enough discriminative power to be used for screening.

Table 2. Comparison of vVectorFp and vectorFp with other fingerprints. The colours show the relative performance of given fingerprint to others on given target (dataset). The grey cell represents the best result on given target while white represents the worst result.



As the first step of the analysis we dropped all the descriptors that were constant which resulted in the elimination of 258 descriptors. In the next step we utilized variance to decide which descriptors have the potential being a useful discriminator. A descriptor taking only two values has a low chance to well discriminate thousands of compounds. As a prestep to variance analysis we had performed normalization on every descriptor. First, we had removed outliers from every descriptor (values outside the second and third quantile), then we normalized the data into the [0, 1] interval using the min-max normalization. After the normalization we computed variance (var_{norm}) for each descriptor. Many descriptors ended up with $var_{norm} = 0$. For example in case of *nAcid* descriptor, about 96.7% of fragments have zero value. This does not leave much space for other values, and basically divides all fragments into just few categories (3 in case of *nAcid*). If we consider the second and third quantile only we get zero variance. This step eliminated 326 descriptors. Since we used these descriptors later in the experiments, we formed group from them

called *constVarQ* (constant variance on quantiles). The remaining 185 descriptors were split into 4 groups of almost the same size based on the value of variance. The groups were called *var_00_25*, *var_25_50*, *var_50_75* and *var_75_100*.

4.2.1. Single descriptor

In the first step we evaluated the performance for selected descriptors from groups *var_00_25*, *var_25_50*, *var_50_75* and *var_75_100* and *constVarQ*. The descriptors in *constVarQ* group performed worst of all, as expected. This was caused by the fact that in many data sets the descriptors were constant and so had no discriminative power. However, despite the overall bad performance few exceptions emerged. For example, using the descriptor *nAcid* (number of acidic groups) for target 20174 resulted in *auc* 0.909. The *tt*, *hashtt* and *ap* scored 0.841, 0.8430 and 0.8450 respectively. This demonstrates that good performance can be reached even with a single simple descriptor. As for the target 20174, the best performance (0.9560) was obtained by *vectorFp*.

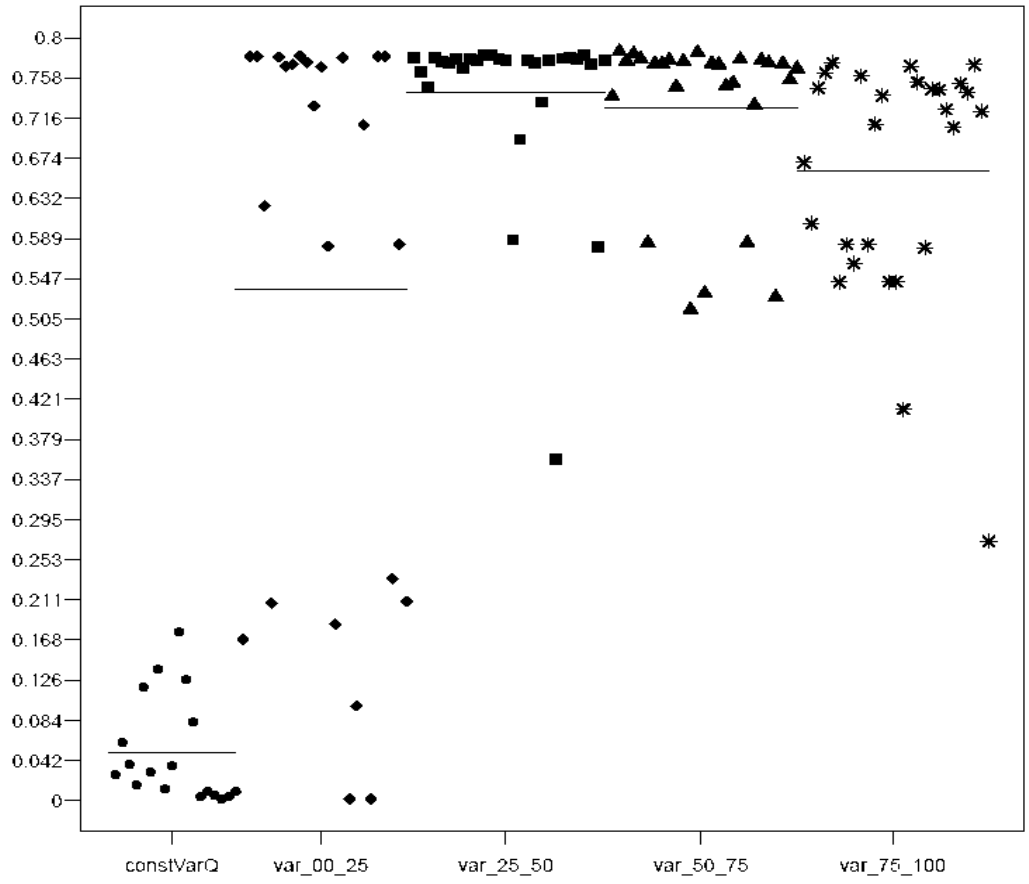


Figure 2. *auc* performance for single descriptor parameterization among the variability groups. The horizontal lines represent the average *auc* for given group.

The performance of all the descriptors shows Figure 2. where descriptors' data points in the same group share same shape. The X-axis corresponds to the individual descriptors while the Y-axis shows the average AUC over all the targets for each of the descriptors. The horizontal line then represents the average of the descriptors performance for each of the group. We can clearly see that the descriptors in the *constVarQ* show worse performance then descriptors in the other

groups. In all the var groups we can identify several well performing descriptors. However, the group *var_00_25* contains many descriptors showing very poor performance. It follows that a descriptor with low variability is likely to perform poorly. On the other hand, the descriptors with high variability (group *var_75_100*) also lead to worse performance than the descriptors with moderate variability (groups *var_25_50* and *var_50_75*).

4.2.2. Multiple descriptors

In the next step, we first created pairs and then triplets of descriptors and used them to label the fragments. Thus, in the previous step each fragment was labelled by exactly one descriptor but in the second step pairs and triplets of descriptors were utilized. Our hope was that the performance would increase when using tuples in contrast to using single descriptors alone.

Since we did not have sufficient computational resources to test every possible pair and triplet of descriptor we implemented a filter. The purpose of the filter is to filter out such tuples which are unlikely to lead to best results. The filter utilizes AUC (*auc*) of single descriptors and correlation (*cor*) between pairs of descriptors.

Let n denote the number of descriptors that should be used in the parameterization (in our case n is 2 or 3). Let auc_i denote the average AUC for *i*-th descriptor (out of n) and $cor_{i,j}$ the correlation between AUC values of *i*-th and *j*-th descriptor over data sets. Thus if two descriptors show similar AUCs over all data sets they have high correlation. In order for the tuple of descriptors to pass the filter, the following two conditions need to be satisfied:

$$\sum_{i=0}^{n} auc_i > auc_{Level}$$

$$cor_{LevelMin} < \max_{0 \le i,j \le n, i \ne j} cor_{i,j} < cor_{LevelMax}$$

The filter is parameterized by the values auc_{Level}, $cor_{LevelMin}$ and $cor_{LevelMax}$. Using auc_{Level} simply prefers tuples consisting of descriptors which behave well when used alone. The idea behind restricting the correlation is that bringing together correlated descriptors would not result in new information and thus probably would not increase the discriminative power of the resulting molecular representation. We tried several parameterizations of the filter (see Table 3.) to get a reasonable number of pairs/triplets for our experiments.

The descriptors in 2B pairs are required to have *cor* between 0.47 and 0.6. The lower bound for *cor* secures that the pairs in 2A and 2B are different. As the trade of, the required *auc* needs to be slightly higher. As the 2B group was selected with less stress on *cor* it was expected that the paired descriptors would have more similar results over the data sets. The goal of different parameterizations of the filter was to test which combination of correlation and quality parameters leads to better results. The same holds for the groups of triplets.

Table 3. Specification of tested filters

group name	auc_{Level}	$cor_{LevelMin}$	$cor_{LevelMax}$	group size
2A	1.48	0.00	0.47	102
2B	1.487	0.47	0.60	40
3A	2.08	0.00	0.50	41
3B	2.20	0.50	0.60	48

How different number of descriptors used to label the fragments influence the *auc* show Table. 3. and Figure 4. The Table 3. average *auc* for parameterizations using single descriptors (the *var* groups), pairs of descriptors (the 2A and 2B group) or triplets of descriptors (the 3A and 3B group) to label the fragments. As the results show the 2A-filtered pairs of descriptors perform on average significantly better those based on the 2B filter.

The difference between 2A and 2B is much higher than in case of 3A and 3B. As Table 4. shows, the performance of triplets of descriptors is somewhere between the 2A and 2B based pairs. From the Figure 3 it seems that the performance of triplets of descriptors is more variable than in case of pairs. We believe that it is the consequence of the fact that there are more triplets of descriptors than there are pairs. Therefore, it is more difficult to identify the correct triplets. Thus the variance in the results of triplets of descriptor is simply due to the imperfection of the descriptor selection procedure. In case of both pairs and triplets of descriptors, the group with more restricted *cor* seems to provide better results, especially in terms of worst case performance.

Table 4. Average reached *auc* for different numbers of descriptors per fragment

group name	average auc
var_00_25	0.535
var_25_50	0.741
var_50_75	0.725
var_75_100	0.659
2A	0.783
2B	0.780
3A	0.782
3B	0.782

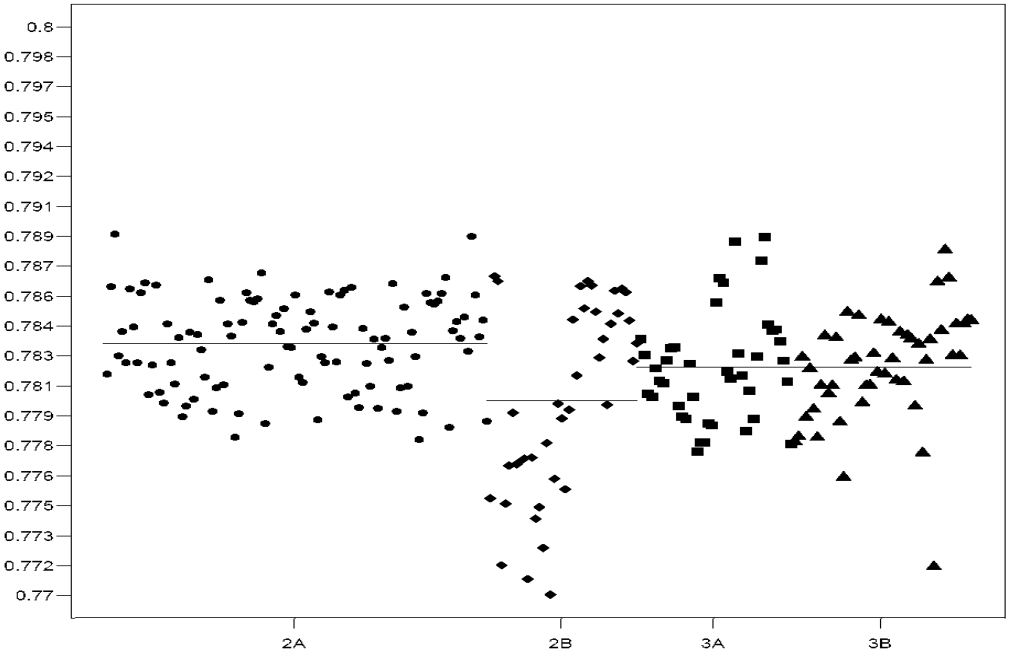


Figure 3. *auc* performance for two and three descriptors parameterizations among groups. The horizontal lines represent average of *auc* for given group.

4. CONCLUSION

In this work we presented a generic molecular representation called VectorFp. The representation was tested using the recently published benchmarking platform for LBVS. Therefore, the results should be easily reproducible and results were easily comparable with other existing commonly used molecular representations. The main motivation for our work was to provide a molecular representation which could be parameterizable with specific descriptors suitable for given biological target. Even though we operated within the boundaries of the benchmarking framework by forcing us to fix the parameterization our method it still outperformed most of the existing methods.

We also showed the potential of our method by creating a virtual representation *vVectorFp* where the best encountered parameterization for given target was used. This representation clearly outperformed all the existing approaches showing that potential strength of the method with correct parameterization. As a virtual representation demonstrates the potential of VectorFp if the right parameterization is used, it follows that the research on the parameterization will be the main direction of our future work on VectorFp. Moreover, we tested only up to three descriptors per parameterization while there are, beside the computer memory, virtually no restrictions of how many descriptors can be used. Finally, we also plan to investigate the possibility of stressing the importance of a single descriptor by its multiple application.

ACKNOWLEDGEMENTS

This work was supported by Grant Agency of of Charles University [project Nr. 154613], by SVV-2014-260100 and by the Czech Science Foundation grant 14-29032P.

REFERENCES

[1] B. Battersby and M. Trau, (2002) "Novel miniaturized systems in high throughput screening," Trends in Biotechnologi, vol. 20, pp. 167–173

[2] J. Besnard, G. F. Ruda, V. Setola, K. Abecassis, R. M. Rodriguiz, X. P. Huang, S. Norval, M. F. Sassano, A. I. Shin, L. A. Webster, F. R. Simeons, L. Stojanovski, A. Prat, N. G. Seidah, D. B. Constam, G. R. Bickerton, K. D. Read, W. C. Wetsel, I. H. Gilbert, B. L. Roth, and A. L. Hopkins, (Dec 2012) "Automated design of ligands to polypharmacological profiles," Nature, vol. 492, no. 7428, pp. 215–220

[3] S. Ekins, J. Mestres, and B. Testa, (2007) "In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling," British Journal of Pharmacology, vol. 152, pp. 9–

[4] R. S. Ferreira, A. Simeonov, A. Jadhav, O. Eidam, B. T. Mott, M. J. Keiser, J. H. McKerrow, D. J. Maloney, J. J. Irwin, and B. K. Shoichet, (Jul 2010) "Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors," J. Med. Chem., vol. 53, no. 13, pp. 4891–4905

[5] L. R. Vidler, P. Filippakopoulos, O. Fedorov, S. Picaud, S. Martin, M. Tomsett, H. Woodward, N. Brown, S. Knapp, and S. Hoelder, (Oct 2013)"Discovery of novel small-molecule inhibitors of BRD4 using structurebased virtual screening," J. Med. Chem., vol. 56, no. 20, pp. 8073–

[6] K. Heikamp and J. Bajorath, (Jan 2013) "The future of virtual compound screening," Chem Biol Drug Des, vol. 81, no. 1, pp. 33–40

[7] O. Taboureau, J. B. Baell, J. Fernandez-Recio, and B. O. Villoutreix, (Jan 2012) "Established and emerging trends in computational drug discovery in the structural genomics era," Chem. Biol., vol. 19, no. 1, pp. 29–41,

[8] P. Ripphausen, B. Nisius, L. Peltason, and J. Bajorath, (Dec 2013) "Quo vadis, virtual screening? a comprehensive survey of prospective applications," Journal of Medicinal Chemistry, vol. 53, no. 24, pp. 8461–8467

[9] P. Ripphausen, D. Stumpfe, and J. Bajorath, (Apr 2012) "Analysis of structure based virtual screening studies and characterization of identified active compounds," Future Med Chem, vol. 4, no. 5, pp. 603–613
[10] W. L. Jorgensen, (Nov 1991) "Rusting of the lock and key model for protein-ligand binding," Science, vol. 254, no. 5034, pp. 954–955
[11] M. A. Johnson and G. M. Maggiora, (1990) Concepts and Applications of Molecular Similarity. Wiley-Interscience
[12] L. Xue and J. Bajorath, (Oct 2000) "Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening," Comb. Chem. High Throughput Screen., vol. 3, no. 5, pp. 363–372
[13] J.-L. Faulon and A. Bender, (Apr 2010) Handbook of Chemoinformatics Algorithms (Chapman & Hall/CRC Mathematical & Computational Biology), 1st ed. Chapman and Hall/CRC
[14] R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, (1987) "Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors," Journal of Chemical Information and Computer Sciences, vol. 27, no. 2, pp. 82–85,
[15] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, (2004) "Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures," Organic and Biomolecular Chemistry, vol. 2, pp. 3256–3266
[16] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service." Journal of Chemical Documentation, vol. 5, no. 2, pp. 107–113
[17] (2014)"Rdkit: Cheminformatics and machine learning softwares," [Online]. Available: http://www.rdkit.org
[18] S. Riniker and G. Landrum, (2013) "Open-source platform to benchmark fingerprints for ligand-based virtual screening," Journal of Cheminformatics, vol. 5, no. 1, p. 26,
[19] (2014) "Python," [Online]. Available: https://www.python.org/
[20] R. P. Sheridan, S. B. Singh, E. M. Fluder, and S. K. Kearsley, (2001) "Protocols for bridging the peptide to nonpeptide gap in topological similarity searches," Journal of Chemical Information and Computer Sciences, vol. 41, no. 5, pp. 1395–1406
[21] J.-F. Truchon and C. I. Bayly, (2007) "Evaluating virtual screening methods: Good and bad metrics for the early recognition problem," Journal of Chemical Information and Modeling, vol. 47, no. 2, pp. 488–508
[22] N. Huang, B. K. Shoichet, and J. J. Irwin, (2006) "Benchmarking sets for molecular docking," Journal of Medicinal Chemistry, vol. 49, no. 23, pp. 6789–6801,
[23] K. Heikamp and J. Bajorath, (2003) "Large-scale similarity search profiling of chembl compound data sets," Journal of Chemical Information and Modeling, vol. 51, no. 8, pp. 1831–1839
[24] S. G. Rohrer and K. Baumann, (2009) "Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data," Journal of Chemical Information and Modeling, vol. 49, no. 2, pp. 169–184,
[25] D. Hoksza and P. Škoda, (2014) "2d pharmacophore query generation," in Bioinformatics Research and Applications, ser. Lecture Notes in Computer Science, M. Basu, Y. Pan, and J. Wang, Eds. Springer International Publishing, vol. 8492, pp. 289–300
[26] C. W. Yap (2011) "Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints," Journal of Computational Chemistry, vol. 32, p. 1466-1477

AUTHORS

Petr Skoda received his master degree in 2014 from Faculty of Mathematics and Physics at the Charles University in Prague. Since 2014 he is Ph.D. studentunder the supervision of David Hoksza. His main research interest is chemoinformatics.

David Hoksza received his Ph.D. in 2010 from the Dept. of Software Engineering, Charles University inPrague. Since 2011 he is an associated professor of software engineering in the department of Software Engineering at the CharlesUniversity in Prague. His research interests include structural bioinformatics, chemoinformatics, data engineering and similarity searching.

USER-CENTRIC PERSONALIZED MULTI-FACET MODEL TRUST IN ONLINE SOCIAL NETWORK

Liu Ban Chieng[1], Manmeet Mahinderjit Singh[1], Zarul Fitri Zaaba[1], Rohail Hassan[2]

[1]School of Computer Sciences, University Sains Malaysia, 11800 Penang
lbchieng@ucom11.cs.usm.my; manmeet@cs.usm.my; zarulfitri@cs.usm.my

[2]Universiti Teknologi Petronas, Tronoh, 31750, Malaysia.
rohail.hassan39@gmail.com

ABSTRACT

Online Social Network (OSN) has become the most popular platform on the Internet that can provide an interesting and creative ways to communicate, sharing and meets with peoples. As OSNs mature, issues regarding proper use of OSNs are also growing. In this research, the challenges of online social networks have been investigated. The current issues in some of the Social Network Sites are being studied and compared. Cyber criminals, malware attacks, physical threat, security and usability and some privacy issues have been recognized as the challenges of the current social networking sites. Trust concerns have been raised and the trustworthiness of social networking sites has been questioned. Currently, the trust in social networks is using the single- faceted approach, which is not well personalized, and doesn't account for the subjective views of trust, according to each user, but only the general trust believes of a group of population. The trust level towards a person cannot be calculated and trust is lack of personalization. From our initial survey, we had found that most people can share their information without any doubts on OSN but they normally do not trust all their friends equally and think there is a need of trust management. We had found mixed opinions in relation to the proposed rating feature in OSNs too. By adopting the idea of multi-faceted trust model, a user-centric model that can personalize the comments/photos in social network with user's customized traits of trust is proposed. This model can probably solve many of the trust issues towards the social networking sites with personalized trust features, in order to keep the postings on social sites confidential and integrity.

KEYWORDS

Online Social Network, Trust, Multi-Faceted Model, Trust Management, Usable Security.

1. INTRODUCTION

Online Social Network (OSN) can be defined as a free online platform, with high availability that serve as a digital representation of the users stay connected in the virtual environment that provide data sharing, semi- public profile creation, and messaging services [1,2,3,4] Online Social Network (OSN) such as Facebook, Twitter and Myspace have experienced a bullet's speedy growth in recent years. Despite the social hierarchy, almost everyone, with an online device, will have at least one account in any of the social network sites. A survey done by [1] has demonstrated that the users of social networking site from 2005 - 2012, consist of people from

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 245–259, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41220

different age group, ranging from 18 to 65 and above. The number of ONSs users has increased in all age groups over the years. The main problem in the current OSN is the generalization of trust in OSN. Friends in a group are assumed to be trusted equally. Take for example, on Facebook and Twitter; they have grouped all friends under one level of the category in which they tend to trust them all the same. Although they can group friends into "Close Friends" and "Family" like in Facebook, the categorization is still in a big group but not personalized and specific. However, in real-life, it is impossible to do so as trustworthiness is context dependent and need to personalized [5, 6]. Some friends are likely to be more trustworthiness compared to the rest. For example, Alice wants to share a private message in the Facebook only with certain friend in the 'Close Friends' group. However, trust level can be varied according to the times, experiences and individual. Moreover, the person you trust before not necessary to be trusted by you in your entire life. As an account owner, have no right calculate your trust level too and the trust level is assumed to be general for all. This research aims to answer the questions of whether a multi-faceted model of trust that is personalisable and specialisable be welcomed in OSNs, would an application of the model satisfy user needs when expressing their subjective views on trust in the OSN environment, and would the proposed solution address issues we found in the literature review. The main aim of this research paper is to tackle the lack of personalization in term of trustworthiness in the current OSN. The objectives of this research paper are stated as below:

1. To study the security, privacy and trust issues in current social network and explore various trust traits and users requirement that is essential to the users.

2. To present the outcome of the questionnaire and to solve the issue of single level trust adopted in the current social network.

This paper is organized as follows; Section Two introduces the concept of OSN, the categorizations as well as a brief history of them. It provides an analysis of the state of the art in trust and its characteristics, and current trust mechanisms used in notable online social networks. Section Three present the gap of the existing online social networks. This chapter also discusses about the trust identification in current OSN. Section Four concentrates on a survey designed to gather user opinions of current trust management approaches being used, and presents our findings as well as analysis of the results and future works. And finally, we come out with a conclusion, discussing the extent to which the original objectives and goals were achieved during this research project.

2. BACKGROUND

In this section, we will present the related section regarding our study such as social network and the trust mechanism in current social network.

2.1. Social Network (OSN)

Online Social Network not only serves as a communication tool but also act as an application source and online community builders [1, 2 3]. Face to face interaction is eliminated in OSNs [4, 7]. It was the most popular internet sites mushrooming in the past few years and today having billions of users with a wide demographic range. Nowadays, the users of OSNs are spread over all age groups despite their backgrounds. The first recognizable OSN is the SixDegrees.com with the initial purpose creates profiles and listing friends in 1997. OSN experienced various evolutions from 1997 till now, with the addition of function, improvement of the interface and the availability of OSN simultaneously with the increment of the popularity of OSNs [8].

2.2. Trust Mechanism in Current OSN

Network in OSNs has become more and more diversify since social sites bring together people, often from different type of social ties; consisting of thick bonding and weak bonding [4]. Hence, forming a different level of trust among "friends" in the social networking sites. "Thick trust" is formed among those sharing common interests in offline interaction while "thin trust" is formed across strangers. He claimed that mixing of social circles in OSN could gradually lead to social distrust [4]. Hence, privacy management should be examined together with the trust model in OSNs. OSNs have been believed to generate many security and privacy issues, and thus, trustworthiness in social networks has been doubted after all. However, trust is an important concept in obtaining the user's heart to use the sites. This is because, a certain level of trust is needed in order to make the user willing to use the sites and share their private data on the sites. The characteristic of trust can be concluded as [1].

i. Trust is asymmetric: Trust is not identical; A might trust B fully but A doesn't necessary to trust with the in the same way.

ii. Trust is transitive: A and B trusts each other well and B has a common friend C, that A might not know where A might trust C because of B. However, A might not trust D, a friend of C since their network linkage is getting far.

iii. Trust is context dependent: In other words, trust level towards an individual can be varied based on time, situation and experience. Depends on the context, people tend to trust each other differently.

iv. Trust is personalized: Which means trust is subjective. Two persons can have different opinions regarding the trust level towards a same person.

Currently, the trust model in social networking adopts the following characteristic:

a) Single- faceted: The current trust model focus only on one trust characteristic, which is an inadequate model of trust since the Internet environment is so broad and the population of users is wide. It is too general in term of trust beliefs and it has ignored a lot of other important trust concepts such as reputation in their model [5, 6]. Dishonesty can happen [5]. However, trust concepts are very useful in considering the relationship between peoples and it should not be unitary but diverse [7].

b) Not personalized: Trust model should be personalized and conjunction with the domain specific model [5, 6]. However, current trust model itself does not inhibit a personalized concept, which take-in consideration of the subjective nature and the views of human's trust towards peoples across a large population [6]. In the real world, trust is context dependent and peoples tend to judge people differently with different weight of trust traits.However, current social networking sites cannot specify the trust level based on the user's customized trust traits on specific individuals.

c) Trust level cannot be annotated or calculated [6]: Friendship is not well- categorized in the current social networking sites [1]. Hence, the trust level towards different individual cannot be explained in context and yet cannot be calculated accordingly [6]. Thus, the trust value on each "friend" is being uniformity with lists or category, but not differentiated according to percentage of trustiness and how the user weighted the importance of trust traits.

2.3 Related Work on Trust Management

The Trust Management Model [5] such as the Marsh's trust model is one of the pioneers to introduce computational concepts of trust and has represented trust in scalar form while SECURE makes it in a range from including the measure of uncertainty. There are also some simple trust calculation in some of the online community like eBay and Amazon, to enable the members to understand the statements and guide them for purchasing and moreover send feedbacks. However, it is based on single-faceted approach and dishonesty can still happen as mostly they will tend to avoid negative comment [5, 6]. Many other trust management systems such as REFEREE, SULTAN, Advogato and Film Trust applied the single- faceted approach, which means they do not inhibit the subjective nature of trust in their users [6].

However, my Trust, Trust Management Service (as shown in Figure 2) has been using a personalized model in conjunction with a specific domain model to provide the user a personalized- trust based services. myTrust has adopted an internal trust multi-faceted management system, TRELLIS, with the trust calculation mechanism based on rating. myTrust has been designed to enable users to annotate trust in term of the traits of trust, share the trust information and the calculate trust. The model is modelled through 4 unique models: Upper Ontology provides generic traits of trust, Meta model, Domain Specific model and personalized model [6].The multi-faceted idea has utilized the subjectivity of trust nature and view found among the large population. The trust concept such as: honesty, reputation, competency, credibility, confidence, reliability, belief and faith are recognized as the core of this multi-faceted model. Besides that, multi-faceted model is able to support personalization and is context dependent. The multi-faceted of trust and the relationship between the trusts concepts are utilized to reflect the subjectivity of human being into the model [6].

Moreover, a multi-faceted management interface that is applicable to both operational and contractual operations [9]. The heterogeneous web services with different levels of capabilities and characteristics can be managed with this multi-faceted interface. There are basically three facets of web services: No management, operational management and contract management. The web services might exist in different domain with different controllers too makes it harder to be manageable. Standard exists for operational management. However, ROAD framework is used to implement the management interface for self-managed mechanism.

Since we have noticed that there are a lack of flexible and personalised trust management features within current OSNs and we believe that such features are important to protect the privacy of users, so we decided to explore whether the multi-faceted model of trust proposed by [6] that enables personalization and the flexibility of annotating trust subjectively would be utilized in OSNs. We also interested to know the desired functionalities of the trust management model to be implemented into OSNs. Based on the research questions, we developed an online questionnaire protocol that included 12 closed-ended questions on four topics: (a) general usage patterns related to content sharing, (b) experiences regarding privacy that are related to content sharing, (c) how and with whom the people share content, and (d) the perceptions of social trust and the desired trust traits requirement.

2.4 A Review on Usable Security in OSN

In view of an increasing threat landscape, today's users face an increasing requirement to use applications, security tools and interact with related system functionality. However, a significant challenge in many cases is posed by the usability of the technologies, with the consequence that users can face difficulties in understanding them correctly and utilizing them effectively. Therefore, Online Social Network (OSN) also facing the similar challenges. Security and usability (Usable Security) are two different domains which become the concern in viewing the

OSN. In order to design usable technologies, it must be designed with a secure applications and interface so that end-users would be able to comprehend the functions provides for them [10]. Nowadays, most researchers have started to focus on this issue and also the privacy and identity albeit users do not want to reveal the information as they wanted [11]. Most of the OSN products provides with security settings which offer from basic to advanced security protection for their end-users. However, most of the end-users would apprehend the default settings which had been provided once they installed the application in their computer rather than having some tweak on the settings for better protection. The main question that can be highlighted here is why the end-users would do that. There are some concerns on how current security settings have been implemented on each application or products which make it cumbersome to be used. By relying upon the default setting is not an adequate solution given the facts that single default level of security unable to serve to all level of users. The developers should not put the end-users in baffle situation where they need to deal with it without a proper guide. In this context, the presentation and usability plays an important role to ensure that end-users able to understand how to manage their security functionality.

Usability can be viewed as a quality attribute that evaluates how user interface is being used [12]. Thus, in order to deliver the information and meet the purpose of one particular system, the end-results of application must be user-friendly and be presented with a correct Human Computer Interaction (HCI) aspects. Many HCI guidelines have been implemented to guide the developers to provide a meaningful manner and understanding system functionality (i.e. interface, security features, etc) [13, 14, 15, 16]. At present, there are not very much focus has been given to improve the usable security of Online Social Network (OSN). As the social networking quickly become popular means of communication, there are corresponding needs to ensure that end-user would be competent and easy to interact and to use the application. However, for this current work, the authors make an early observation to compare six types of Online Social Network (OSN) in respective of usability as a template filled with useful template data in a general sense as shown in Table 1. It is expected that the developers able to re-asses this template to fill in with more detail and concrete data via fully implementation of usability evaluation methods which also become our future works.

Table 1. General analysis of popular social Network based on usability criteria.

OSN Types	**Effectiveness**	**Efficiency through Guiding Interface**	**User Satisfaction**
Facebook	High if the user takes some time to learn (i.e. technical explanation and terminology in two separate tabs)	Guiding user through question and answers with detail explanation which includes different signal cues	Easy for advanced users as it involved some technical terminology. Novice users might face some difficulties to understand (Depend on the usefulness and clarity of questions)
Twitter	Medium (i.e. simple explanation in combination security & privacy tab)	Guiding user through useful tabs with basic explanation	Simple and easy to understand as the security features is basic (Limited choices)
YouTube	Medium (i.e. simple explanation in privacy tab)	Guiding user through simple tabs with limited explanation	Limited information and explanation. However it is easy for all level of users.(Limited choices)
Google+	High if the user takes some time to learn (i.e.	Guiding user through	Long list of information provided. More suitable for advanced user

	technical explanation and terminology in Privacy tab)	long list of information.	rather than novice (Learnability will take some time)
Pixnet.net	Medium (i.e. basic explanation in settings)	Guiding user through useful tabs and help queries function.	Easy for advanced users as it involved some technical terminology. Novice users might face some difficulties to understand (Depend on the usefulness and clarity of questions)
Myspace	Medium (i.e. simple explanation in privacy tab)	Guiding user through useful tabs with limited explanation	The security tab listed in a small font at the bottom with a very limited information and explanation (Problem with visibility)

This study using a similar approach to compare user help techniques based on usability as a new platform by providing some general ideas on the evaluation of each technique [17]. All of these will be evaluated using Usability evaluation Methods (UEM) via inspection, user testing or inquiry. However, with this particular study, the authors made an early observation and general comparison rather than evaluate it in further details. It is expected that the software designer will be able to know elements that can be enhanced or formulated to make their application works better. Usability includes three main attributes such as effectiveness, efficiency and user satisfaction [18]. Effectiveness was measured by looking at whether user able to understand the usage of security settings provided in timely manner. Efficiency focused on whether user able to complete the tasks after learning how to use one particular application while user satisfaction focused on whether the application that user used is pleasant to gain full satisfaction.

3. SURVEY FINDINGS

To gain insight into different practices regarding trust in Online Social Network, the questionnaire groups participants into three categories as follows, people who are currently using OSNs, people who have used OSNs before but are no longer active, and people who have never signed up in any OSNs. In total, 213 people took part in answering the questionnaire. However, only about 200 samples are taken due to the validity and completeness of the survey result return from each participant. Among which, 117 were female, 83 were male. Mostly from age ranges from 23 to 25. Among all the 200 participants that contributed in this survey, there are 179 active OSN users, 12 people that are no longer active in OSNs and 9 people who never or will not sign up in any online social networks.

Among the 179 respondents who are currently using OSNs, the majority of the profiles are set to be viewable by the friends that are directly linked through the users' networks. This indicated that the OSNs users are more comfortable to share their data to people that they know than exposed everything to the public. We then asked the question of whether these users are happy with the available ways of controlling access to their profiles. We found that most people are pleased with current access control methods; they can share their photos and other contents without doubts. Most of them think that the settings are automated with the previous settings and are easy to control too. Most of the users also think that their privacy is protected in OSN and feel safe when using current OSN. Similarly, most of the users believe that OSN will not use their information for other purpose. They feel safe using OSN for content sharing.

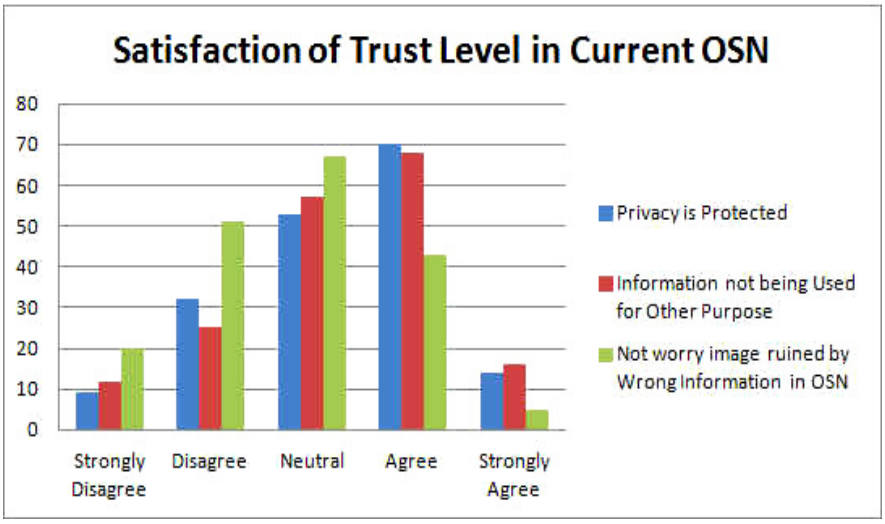


Figure 1. User satisfaction towards current access control methods- Category One

Most of the users also think that their privacy is protected in OSN and feel safe when using current OSN as indicated in Figure 1. There are only less than 40 of the peoples that always doubt about their privacy in OSNs. Also, there are about 55 of people standing neutral instead of addressing concerned in it. Similarly, Figure 1. also indicated that, most of the users believe that OSN will not use their information for other purpose. They feel safe using OSN for content sharing. Only less than 40 of the participants think that it is not, while around 58 of people didn't point out their opinions but remain neutral. As Figure 1. implies also, despite relying too much on OSNs, most of the people are worried about their image is being ruined by wrong information posted in OSN, while about 68 of the respondents stand neutral for it. Only approximately 40 of people are not worrying about it. Since most of the people are satisfied with the current access control methods, we asked the question of whether they trust random strangers to view their profiles, as well as the question of whether access control really is necessary. The result has indicated that only 10 out of these people actually stated the fact that indeed, they do trust anyone and everyone, including random strangers, viewing their profiles. Most people, however, claimed that they do not, while also a small portion of people are not bothered by it at the same time.

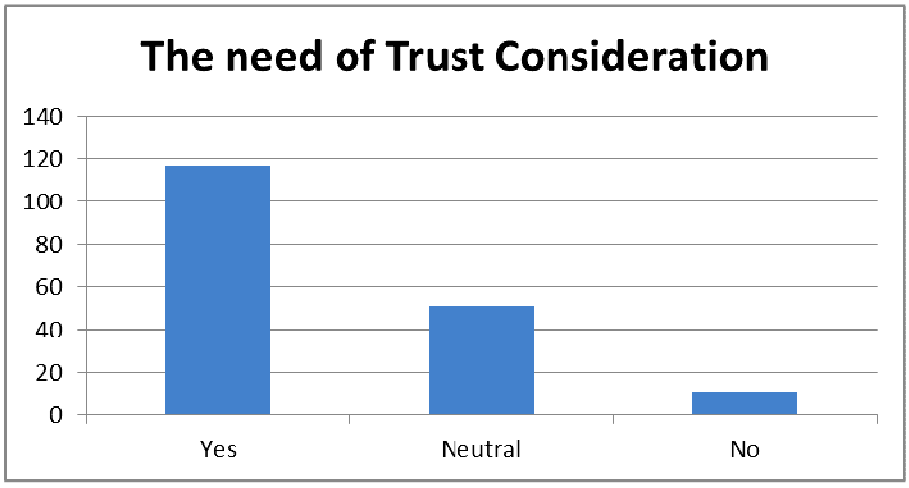


Figure 2. Necessity of access control in OSNs - Category One

We have found a similar contradictive response regarding the necessity of considering the trust level when sharing contents in OSNs, as shown in Figure 2, only less than 10 of these people think it is not necessary, while most people, nearly 120 of the respondents believe that considering the trust level in OSN when sharing something is necessary, and around 50 of people do not care about having control over their profiles and remain neutral. During their memberships of the 12 respondents who are no longer active in OSN, 9 of the participants had set their profiles accessible by directly linked networks, while only less than 1 of them allowed friends of a friend to access their profile. There are only another one of them that set their accessibility to anyone or searchable by search engine. When asked about why you have stopped using OSNs, for instance, a lot of people lost interest in OSNs, mostly due to they are not really happy with the access settings. In our survey, approximately 5 of the participants in this category have lost trust on OSNs most probably due to some unpleasant experiences during their membership. There are around 4 of them who don't dare to post their private data online, as they are doubt for the confidentiality of their data. When asked whether they think access controls of profiles are necessary in OSNs, this group of people had a similar response to category one. On the other hand, among 9 respondents that never signed up in any OSN, some had no interest, some dislike the idea of having private information on the Internet and none of them have never heard of OSNs.

3.1. Desired Trust Features and Opinions on the Proposed Solutions

In contrast, when we asked the 200 people the question whether they would trust all their directly linked friends to view all parts of their profiles, most of the respondents only trust some of their connected friends but not all. Most of the people also feel safe when sharing content but only applied to sometimes, while about 30 of them are doubt about the data confidentially and only less than 20 of them feel definitely comfortable on content sharing. There are only about 10 of them who don't really care about it. We have found a similar contradictive response regarding the necessity of considering the trust level when sharing contents in OSNs, only less than 10 of these people think it is not necessary, while most people, nearly 120 of the respondents believe that considering the trust level in OSN when sharing something is necessary, and around 50 of them do not care about having control over their profiles and remain neutral. We would like to find out if a multi-faceted model of trust that calculates a weighted average of the eight trusts attributes: credibility, honesty, reliability, reputation, competency, belief, faith and confidence, is to be integrated into OSNs, would that be welcomed?

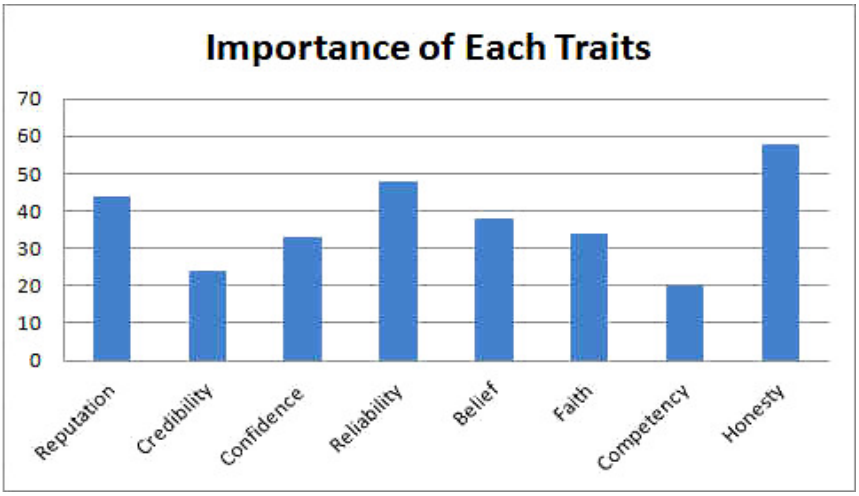


Figure 3. Importance of the 8 Trust Traits

We would like to know is ranking of the eight traits can represent the subjective views of trust in OSNs as well. To do so, we have asked 200 participants who of those eight attributes of trust are most important in their opinions, as shown in Figure 3. honesty appears to be the most important factor, closely followed by reliability and reputation as well as credibility. Many of them think that rating friends in OSNs seem cruel. However, since privacy is an issue they are willing to take the chance, if there is such a setting.

3.2 Analysis of Survey

Several issues have been discovered during the survey, as discussed within this section:

i.Current trust mechanisms need to be refined.

We find out that, in current OSNs, a single faceted mechanism is used, where user can selectively set their profiles accessibility to anyone or specified groups. Even though the users trust each member in a specific group differently, they are not able to state the trust level for each friend separately. Although mostly they are satisfied with the current access settings in OSNs, a large number of people are worried for wrong information spread through OSNs about them. There should be a multi-faceted mechanism that allows users to express their various degrees of trust in a person, or a group of people context-specifically since the main problem in the current system is that, users cannot express their subjective views on trust freely, and the fundamental trust characteristics mentioned in section 2.4 are not utilized in OSNs.

ii.Need of better control on the accessed of profiles

As our findings have contradictive found that, a large number of users do not trust anyone and everyone to view all parts of their profiles, and believe controls are indeed necessary in OSNs. This means that, the existing trust mechanism in OSNs has not achieved user satisfaction, hence, refinement of trust management is needed in OSNs.

i.Users are unsure about a multi-faceted model of trust with rating features.

Other contradictive findings in this survey are that, users think that trust level should be refined in OSNs, but on the other hand, users have not agreed with the rating features. They find it hard to rate someone they know personally and been rated by others too. Such opinions could be the result of a lack of understanding regarding the proposed solution, as for a large percentage of candidates, since the word rating is so open to be interpreted, it would be very hard for them to simply imagine what ratings could be like without having the rough ideas of how it is going on.

4. SYSTEM DESIGN AND IMPLEMENTATION

With influences from Quinn's trust model and considerations for user requirements, we introduce miniOSN, an online social network with a trust rating feature implemented. miniOSN is a web based system that has functionalities of a basic online social networking website, it allows users to create accounts for themselves with a username and password. Users of miniOSN can then confirm a friend request, edit their friendships, upload photos, post status, as well as edit the trust requirements for their content shared.

MiniOSN is developed using PHP programming language and hosted in miniosn.comyr.com, as a free web site. By using miniOSN, users can view a list of his/her friends and edit the friendship accordingly. As illustrated in Figure 4, the user can specify the trust towards his/her friend. The user can rate each of the connected friends differently according to the trust between them based on the eight trust traits, namely honesty, reliability, reputation, credibility, confidence, competency, faith and belief. The rating is made by default zero value, and is range from 1-10. The larger the number, the higher the rating it is.

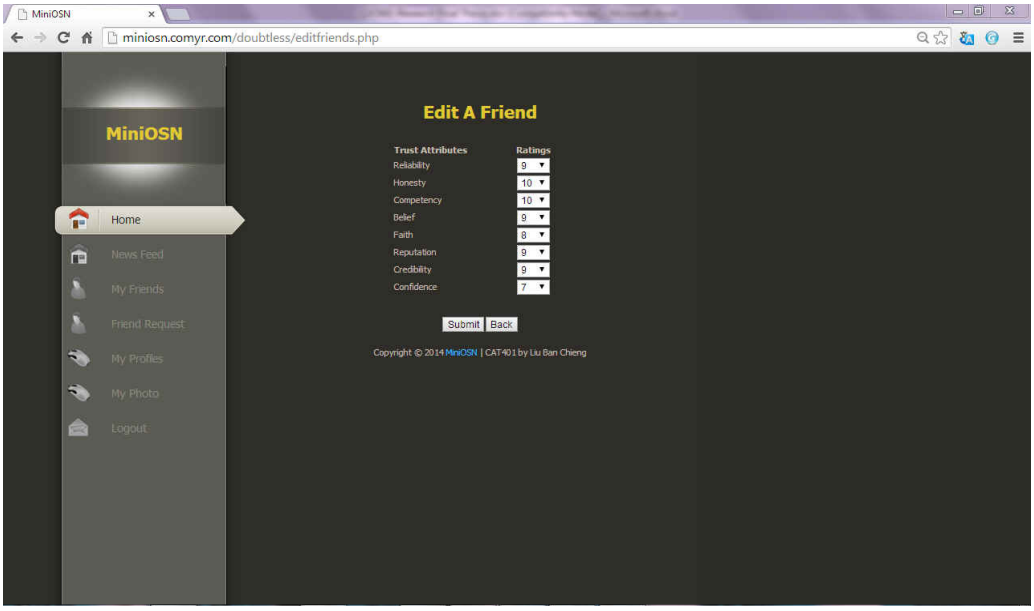


Figure 4. Edit Friendship

MiniOSN enables users to post a status and a photo. Figure 5. shows the timeline of the user profile if the user clicks “My Profile”. User is able to post a status under “What’s On Your Mind?” At the same time, the user can set the weight of each trust traits accordingly in order to adjust the accessibility of the post he/ she want to share.

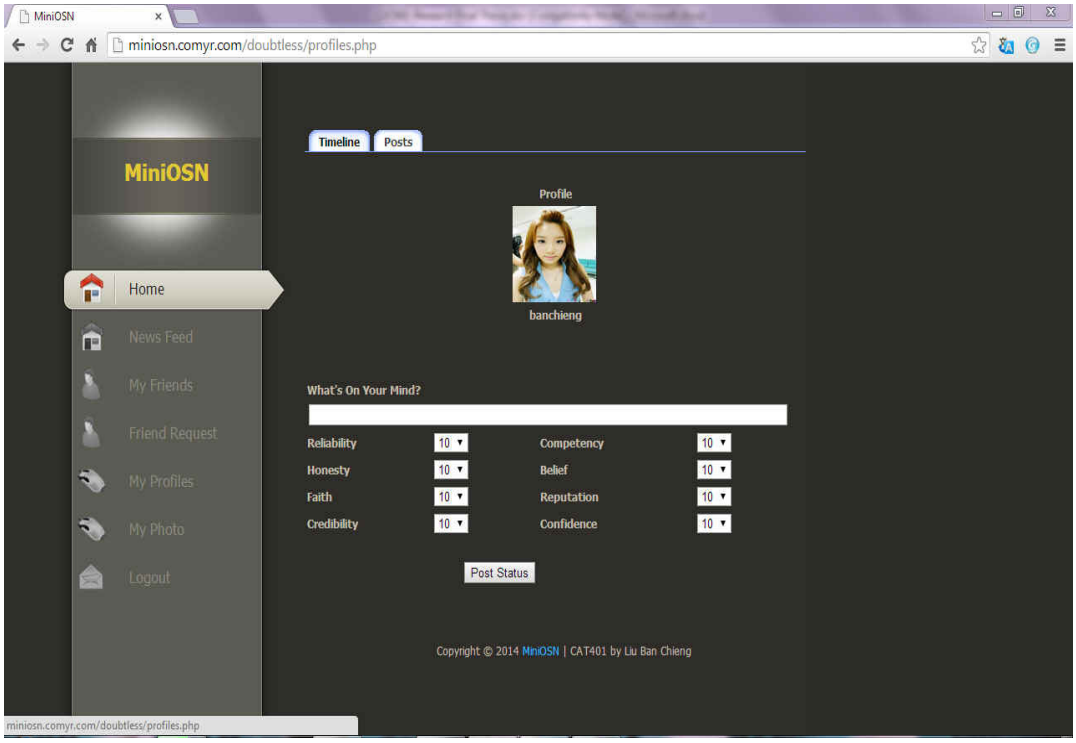


Figure 5. Profile Page

5. EVALUATION

In order to find out what users think of the design and functionalities of miniOSN in relation to expressing various subjective views on trust, we chose to conduct another survey targeting the active OSN users from the 1st survey and gather their opinions on the proposed solution.
Our purpose of the evaluation survey is to find out:

- Whether the user accept the idea of expressing various degrees of trust among connected friends
- How well the trust characteristic found in the literature review helps users of minors to express trust towards friends
- Is the proposed rating feature helpful in gaining better control of user profiles and the content shared?
- The limitation and weaknesses of the mini OSN.

Similar to the 1st survey method, as we are still aiming at a large audience, therefore, Google Docs was again chosen to host the survey on the 5th of May, 2014, over a period of one month time. Invitations to take part in the survey were sent out through email and private message in Facebook. However, this time, we are targeting on the active OSN participants from the previous survey to answer our questionnaire.

There are three parts of the survey questions, the first part aimed to find out whether the system meets the functional requirement. The second part is to find out do users feel the need to express their various levels of trust among their connected friends and the participants' opinions on how they felt about the usage of proposed trust management solution and how well can users in mini OSN express their subjective views of trust personally and context-dependently. And finally, we asked participants on how well they understand about the 8 trust traits we implemented in mini OSN and how they felt about the trust rating feature. From all this aspects, we are able to defined possible refinement of miniOSN.

5.1.Evaluation Results

From the first part of the questionnaire, all of the respondents are satisfied with the functions implemented where their content shared which included photos and status are only visible to the trusted friends only. All of the respondents agree that miniOSN works properly without any technical issues. In the second part of the questionnaire, we found that when asked whether participants felt that they could express trust transitively depending on the context, most of the candidates felt that this is indeed the case. Although one of them felt that miniOSN is not modeled well and doesn't help much in expressing trust depending on context and another one of them felt that there is not much different with the existing OSN. However, almost all of them felt that miniOSN help them to express trust personally with only about 20 of them felt that there is no much difference with the existing OSN.

In the third part of the questionnaire, we found that when asked whether participants felt that rating is an ideal way to preset trust between human, most of the candidates felt that this is indeed the case. Although there are also 38 of them who felt that rating is not a good way to express trust and 20 people remains neutral in this case. However, we found that there is a contradictive view as shown in Figure 6, regarding whether to set the rating to be visible to others or just the user. Surprisingly, among 57 of the participants felt that, rating should be made visible to others while 45 of them think in the other ways and a few of them remains neutral or doesn't pose any concern on it.

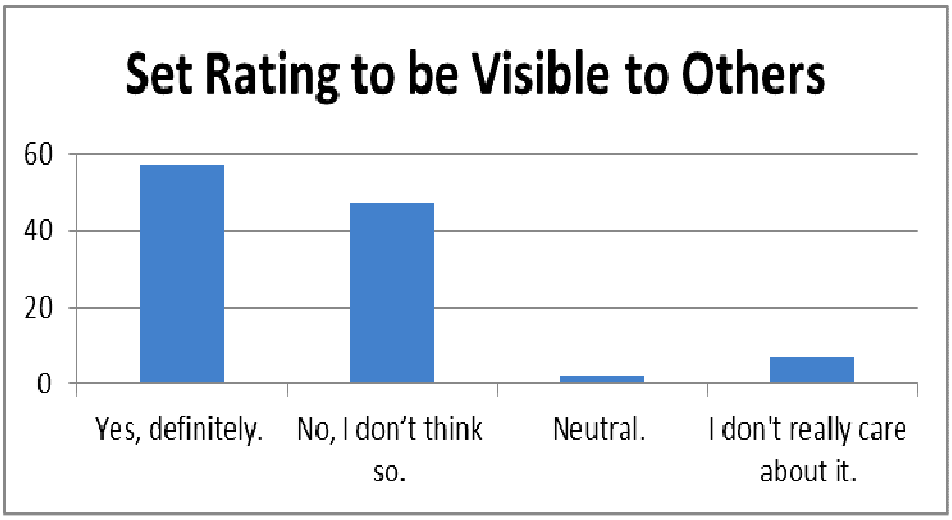


Figure 6. Rating should be Set Visible to Others

When issuing some questions regarding the usage reliability and satisfaction of the trust mechanism, most of the participants felt that miniOSN indeed helps them to shares without any doubts anymore. They felt that their privacy of private information is protected in miniOSN where the weight settings features with preview are convenient for them as it is automated with the previous settings too. However, there is also about 10 of them who do not agree with that, where they found that it is hard to shares without any doubts and they didn't find that the settings helps much. Most of the candidates indicated that miniOSN is easy to use and the rating feature is very helpful in the sense of restrict accessibilities of content shared. However, there are less than 21 out of them think that it is hard to use miniOSN and the rating feature is complicated. Majority of the participants will prefer using miniOSN in the future frequently. In contrast, only two of them will never use miniOSN again. Similarly, all of the participants are confident to use the websites without any doubts and said to be can imagined that most people can use the miniOSN very quickly.

Besides that, we would like to explore the user-friendliness of miniOSN. Figure.7 indicated that about half of the participants think that miniOSN is unnecessarily complex in the sense of the structure and mechanism of rating. They felt that the system become unnecessarily complex and should be modeled in a more simple way that at the same time captured trust mechanism effectively. Only about 30 of them think that it didn't create any difficulties for them in term of complexity. We also found that there is a contradictive view regarding the need of technical support. Majority of the participants felt that they can use the system without any technical assistant, probably because the instruction of using the system are already stated in the homepage that can lead them to use the system easily. Only 22 of them felt that they still need a technical support to assist them in using the system. Moreover, most of the candidates felt that miniOSN is not cumbersome to use and is still considered convenient for them although the structure itself is complex. Only about 10 of the participants felt that miniOSN is not well organized which make it not user-friendly. They might think that the work load has made the simple posting become complicated. When asked whether they need to learn a lot before they can use miniOSN, the responses appear to be dispersed. About half of the total of them felt that it is indeed, while the other half of the participants felt the other way.

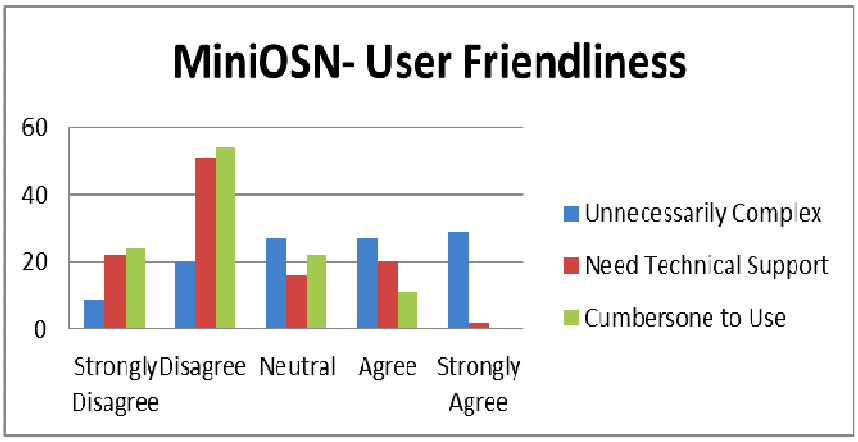


Figure 7. MiniOSN - User Friendliness

Overall, more than half of the participants felt that the trust mechanism implemented in miniOSN did help users to express various degrees of trust, and it also helped users to gain a better control over their resources in online profiles. However, it was mentioned that the rating system in miniOSN seemed to be over-complicating the situation earlier, most probably because some of them found that it is difficult to understand the attributes. Again, they might think that the work load has made the simple posting become complicated. However, overall, most of them think that the concept is okay for them and is easily understood.

5.2 Evaluation Analysis

From the evaluation results, we have found that most people would like to express their subjective views of trust among connected friends depending on the context in the OSN. Most of them felt that the proposed solution would help users to gain a better control over the resources in online profiles. However, some enhancement and modification should be done especially on the structure and design of miniOSN. Since the trust traits concept might cause some confusion and misunderstanding to the users, the selection of trust traits should be defined. Besides that, the management problem is also crucial. Although the users might have a full control over the trust settings of their connected friends, which works well on a one-to-one basis, however, when one has to manage a large number of friends, it becomes difficult for the user to keep track of various sets of numbers.

The design and structure of the miniOSN should also be simpler in the sense that it doesn't complicate the usage of social networks. miniOSN do allowed users to clearly see a list of all their connected friends and their given trust ratings, for easy comparison and readjusting. However, the trust traits number controlling the accessibility should be reduced in the sense to reduce confusion and complication of the overall system. The proposed solution addresses the problem of a lack of personalization when modeling trust in OSNs; however, a common view that trust level decreases as the link between nodes grow longer is not being captured well at the moment. Also, the major problem to be solved is indeed the unnecessary complicated structure of trust mechanism design.

6. CONCLUSION

This paper discussed about the challenges faced in online social networks nowadays. Research has proved that the current issues can be classified into security and privacy which can give a negative impact on the trustworthiness and integrity of social networking sites. The security

impacts include cybercriminals, identity theft and social phishing, stealing the information of the users affecting data integrity and confidentiality [19, 20]. Malware attack could harm the data availability while physical threat can harm the user's life or reputation. Third party application and advertisement can dig user's data through the social networking sites API. Leakage of data can lead to privacy threat such as identity, user and data privacy. Current trust model in social networking sites using the single- faceted approach is said to be not well differentiate the categorization of friends and the trust value is not personalized and specified. Throughout the comparison and contrasts, a multi-faceted model of trust is proposed by adopting the idea from [6].

Based on the outcome of this research, we have extended our work. We have designed a miniOSN, where the trust concerns are taken based on the eight important traits: honesty, reputation, competency, credibility, confidence, reliability, belief and faith [6]. This model is user-centric, personalized and context dependent, which believes can fit the entire trust requirement of the users. On the other hand, we intend to investigate further the assessment on usability elements via effectiveness, efficiency and user satisfaction. We have conducted a second survey based on Usability evaluation Methods (UEM). From the evaluation result, MiniOSN is said to provide users with better control over their online resource but refinement is indeed needed to reduce the complexness of the concept which support our hypothesis by using a multi-facet trust model for media social such as Facebook.

REFERENCES

[1] Johnson.H, Lavesson.N, Zhao.H & Wu.S.F (2011), "On the Concept of Trust in Online Social Networks", Trustworthy Internet, Springer Milan PP. 143-157, [view online] http://link.springer.com /chapter/10.1007/978-88-470-1818-1_11#.

[2] Zhang.C, Sun.J, Zhu.X & Fang.Y (2010) "Privacy and security for online social networks: challenges and opportunities," Network, IEEE, vol.24, no.4, PP.13.

[3] Gunatilaka. D (2011), "A Survey of Privacy and Security Issues in Social Networks." Retrieved From: http://www.cs.wustl.edu/~jain/cse571-11/ftp/social/index.html

[4] Brandtzæg. P.B, Lüdersa.M & Skjetnea. J.H. (2010), "Too Many Facebook "Friends"? Content Sharing and Sociability Versus the Need for Privacy in Social Network Sites," International Journal of Human-Computer Interaction, 26:11-12, PP.1006-1030.

[5] Ruohomaa, Sini, and Lea Kutvonen. "Trust management survey." Trust Management. Springer Berlin Heidelberg, 2005. pp 77-92.

[6] Quinn, Karl. A Multi-faceted Model of Trust that is Personalisable and Specialisable. Diss. School of Computer Science & Statistics, Trinity College, Dublin, 2007.

[7] Chen.X & Shi.S (2009), "A Literature Review of Privacy Research on Social Network Sites," Multimedia Information Networking and Security, 2009. MINES '09. International Conference, vol.1, PP.93-97, 18-20 Nov. 2009.

[8] Boyd.D.M. & Ellison.N.B (2008), "Social Network Sites: Definition, History,and Scholarship," International Communication Association, Journal of Computer-Mediated Communication 13, PP. 210-230.

[9] King, Justin, and Alan Colman. "A Multi Faceted Management Interface for Web Services." Software Engineering Conference, 2009. ASWEC'09. Australian. IEEE, 2009.

[10] Smetters, D. K. & Grinter, R. E. (2002) "Moving from the design of usable security technologies to the design of useful secure applications", Proceedings of the 2002 workshop on New security paradigms. Virginia Beach, Virginia: ACM, PP. 82-89.

[11] Hart, J., Ridley, C., Taher, F., Sas, C. & Dix, A. (2008). "Exploring the Facebook Experience: A New Approach to Usability," Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building, Bridges, ACM. PP. 471-474.

[12] Nielsen, J. (2003) "Usability 101: Introduction to Usability". Retrieved from: http://www.useit.com/alertbox/20030825.html.

[13] Johnston, J., Eloff, J.H.P & Labuschagne, L. (2003) "Security and human computer interfaces", Computers & Security, Vol. 22, no 8, PP. 675-684.

[14] Nielsen, J. (1994) "Ten Usability heuristics". Retrieved from: http://www.nngroup.com/articles/ten-usability-heuristics/.
[15] Norman, N. (2014) "The first principles of human computer interaction". Retrieved from: http://asktog.com/atc/principles-of-interaction-design/
[16] Shneiderman, B (1998) "Designing the user interface: strategies for effective human-computer interaction, Menlo Park, CA: Addison Wesley
[17] Herzog A. and Shahmehri, N. (2007) "User help techniques for usable security", Proceedings of the 2007 symposium on Computer human interaction for the management of information technology, March 30-31, Cambridge, Massachusetts
[18] Hertzum, M. (2010). "Images of Usability" International Journal of Human-Computer Interaction, 26(6), PP.567-600.
[19] Sophos (2011). Social Networking Security Threats: Understand Facebook security threats. Retrieved From: http://www.sophos.com/en-us/security-news-trends/security-trends/social-networking-security-threats/facebook.aspx
[20] Brandt. A. (2010). Five Reasons You Should Always "Stop. Think. Connect." Retrieved from: http://www.webroot.com/blog/2010/10/04/five-reasons-you-should-always-stop-think-connect 2/

INTENTIONAL BLANK

DEVELOPING AN ARABIC PLAGIARISM DETECTION CORPUS

Muazzam Ahmed Siddiqui[1], Imtiaz Hussain Khan[2],
Kamal Mansoor Jambi[2], Salma Omar Elhaj[1,] Abobakr Bagais[2]

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
[2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
maasiddiqui@kau.edu.sa, ihkhan@kau.edu.sa, kjambi@kau.edu.sa, salma53ster@gmail.com, power_baker@hotmail.com

ABSTRACT

A corpus is a collection of documents. It is a valuable resource in linguistics research to perform statistical analysis and testing hypothesis for different linguistic rules. An annotated corpus consists of documents or entities annotated with some task related labels such as part of speech tags, sentiment etc One such task is plagiarism detection that seeks to identify if a given document is plagiarized or not. This paper describes our efforts to build a plagiarism detection corpus for Arabic. The corpus consists of about 350 plagiarized – source document pairs and more than 250 documents where no plagiarism was found. The plagiarized documents consists of students submitted assignments. For each of the plagiarized documents, the source document was located from the Web and downloaded for further investigation. We report corpus statistics including number of documents, number of sentences and number of tokens for each of the plagiarized and source categories.

KEYWORDS

Plagiarism detection, corpus linguistics, Arabic natural language processing, text mining

1. INTRODUCTION

In the academic community, the term plagiarism (synonymous of cheating) is commonly used when someone uses the work of another person without proper acknowledgement to the original source. The plagiarism problem poses serious threats to academic integrity and with the advent of the Web, manual detection of plagiarism has become almost impossible. Over past two decades, automatic plagiarism detection has received significant attention in developing small- to large-scale plagiarism detection systems as a possible countermeasure. Given a text document, the task of a plagiarism detection system is to find if the document is copied, partially or fully from other documents from the Web or any other repository of documents. At a broader level, the researchers have used both extrinsic and intrinsic approaches in developing such systems. The extrinsic plagiarism detection uses different techniques to find similarities among a suspicious document and a reference collection [1], [2], [3]. On the other hand, in intrinsic plagiarism detection, the suspicious document is analyzed using different techniques in isolation, without taking a reference collection into account [4], [5]. Recently, evaluation in plagiarism detection systems has seen considerable attention. One limitation which exist in bulk is the lack of standardized corpus which contains different levels plagiarism, e.g. exact copy, minor

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 261–269, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41221

paraphrasing, extensive paraphrasing and so on. The problem is even worse when we develop and evaluate a plagiarism detection system for Arabic language. This is because research in Arabic natural language processing is still in infancy and we are not aware of any sizeable corpus of plagiarized documents.

In this paper, we present an ongoing research on developing an Arabic plagiarism detection corpus. The need of such corpus is driven by necessity and is two-fold. First, we intend to use this corpus to inform the design of plagiarism detection system. Second, the corpus will serve as a gold standard for automatic evaluation of the proposed plagiarism detection system. Our corpus development approach is closely related to [6] in spirit, but it differed at least in two different ways. First, we develop the corpus for Arabic language whereas [6] built corpus for English. Second, they simulated plagiarism cases in their corpus asking participants to reuse information from other documents intentionally. We collected students samples without explicitly asking them to reuse information from other sources thereby providing genuine cases of plagiarism (details follow).

2. RELATED WORK

There are different methods to build a plagiarism corpus, ranging from collecting genuine examples of plagiarism, or creating a corpus automatically by asking authors/contributors to intentionally reuse another document. This section will provide a representative summary of some of these methods that have been employed to create corpora for plagiarism detection or related topics.

One such example of creating a corpus automatically was presented by [4]. They manually *plagiarized* articles from the ACM computer science digital library by inserting copied as well as rephrased parts from other articles. The purpose was to build a corpus for internal plagiarism detection.

A similar example of an automatically created corpus is the corpus for the 2009 PAN Plagiarism Detection Competition [7]. It simulates plagiarism by inserting a wide variety of text from one set of documents to others. The reuse is either made by randomly moving words or replacing them with a related lexical item or translated from a Spanish or German source document. Similar approach was taken by [8] by inserting a section of text written by different author into a document without changing it.

The METER corpus [9] was manually annotated with three different levels of text reuse: verbatim, rewrite and new. The corpus consists of news stories collected during a 12 month period between 1999 and 2000 in law and show business domains.

To identify paraphrasing, a subtle form of plagiarism, [10] built a corpus from different translations of the same text. The corpus created by [10], along with two other corpora, was manually annotated for paraphrases by [11].

Automatically creating a corpus through text reuse is somehow convenient in the sense that it allows for creation of corpora with little effort. Its disadvantage is that it does not reflect different types of plagiarism that might be found in an academic environment. The corpus created by [6], simulates plagiarism in an academic setting by asking students to intentionally reuse parts of documents in their answers. Our approach is similar to theirs but, in our case, the students were encouraged to use the Web for their research, but were not explicitly asked to plagiarize.

3. CORPUS CREATION

Our original collection consisted of more than 1600 documents in Arabic. More than 1100 of these documents came from the assignments submitted by the students in a first year course about introduction to computers, at our university. In the later part of this paper, we will refer to this set as suspicious documents. The rest were source documents that were located against the suspicious documents and downloaded from the Web. In the later part of this paper, we will refer to this set as source documents. There are several reasons to choose the aforementioned course.

1. The course is offered in Arabic as opposed to the rest of the curriculum, which is in English.
2. It is a mandatory course for every student in the university, which made it possible to collect a large sample.
3. The course is offered by our faculty, which made it easy to collect the data. Our previous efforts to contact other faculties to provide us with students' samples were unsuccessful.

The students were asked to write an essay about the importance of information technology and were encouraged to use the Internet and cite their sources, especially in the case of a website. Since the students were not specifically instructed to copy verbatim or rephrase, different levels of plagiarism exists in the corpus, such as exact copy, light modification or heavy modification.

To get the source documents, references were manually extracted from the suspicious documents. These references were stored with the names and IDs of the suspicious documents. Table 1 displays the basic descriptive statistics regarding the number of references per document.

Table 1: Descriptive statistics about the number of references per document

Statistic	**Value**
Mean	1.46
Median	1
Mode	1
Standard Deviation	1.49
Minimum	1
Maximum	21

The distribution of number of references per document is given by. Most of the documents contain only one reference with the exception of one document containing 21 references, which is evident from the histogram. The document was manually inspected to verify if the outlier was caused by a document processing error or if it was a real value.

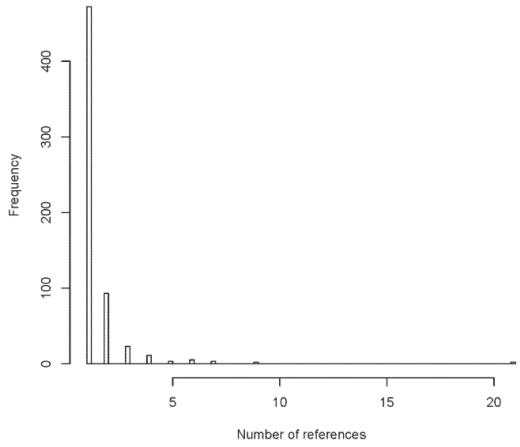


Figure 1: Distribution of the number of references per document

For the suspicious documents where the source URLs were provided, the source documents were located and downloaded from the Internet. To download the source documents, we used a crawler that, given the list of source URLs, downloaded the HTML pages. The pages were cleaned of the HTML tags and the text was extracted from each page. The crawler was written in Java and the text processing was done in Python. The resulting documents were saved in text format with a reference to their sources to identify a suspicious – source document pair.

4. CORPUS ANALYSIS

The corpus was analyzed to compute the basic descriptive statistics. This section will provide statistics including the plagiarism related statistics and sentence and token level statistics from the corpus. Gathering the latter two is important, especially for computing measures for intrinsic plagiarism detection.

4.1. Corpus Statistics

As discussed above, our corpus consists of assignments submitted by the students in one course. Most of these submitted assignments were in MS Word format, but some were in PDF or other formats too. We converted the submitted assignments to plain text format for further processing. This resulted in some processing errors where we were not able to convert a particular suspicious document to the text format. The corpus statistics after text processing and cleanup will be described in the later part of this paper. Different types of statistics were gathered from the corpus. These include the plagiarism related statistics and sentence and token level statistics. The latter two are especially important in building an intrinsic plagiarism detection system.

4.1.1. Plagiarism Statistics

For the purpose of corpus building, the suspicious documents where the references were provided were considered as plagiarized. Documents where the reference was not provided were manually analyzed for plagiarism. The provided reference was used as a label identifying the document as plagiarized and, in case, if the reference contains one or more URLs, the source documents were fetched from the web create a suspicious – source document pair. Some of the documents were plagiarized from the web but instead of providing a URL, terms such as ويكيبيديا (Wikipedia), الانترنت (the internet) were given as a reference. Table 2 displays the plagiarism related statistics.

Table 2: Corpus statistics before cleanup

Type	Count	Proportion
Total number of documents in the corpus	1665	
Total number of suspicious documents	1156	69.4% of total
Total number of source documents	509	30.6% of total
Plagiarized documents	892	77.2% of suspicious
Not plagiarized documents	264	22.8% of suspicious
Documents plagiarized from the web	718	80.5% of plagiarized
Documents plagiarized from other sources	174	19.5% of plagiarized
Documents plagiarized from the web with source URL provided	551	76.7% of web plagiarized
Documents plagiarized from the web without source URL provided	167	23.3% of web plagiarized

4.1.2. Sentence Statistics

For sentence segmentation in colloquial Arabic [12] provided simple heuristics to identify sentence boundaries. These included the use of punctuation marks and newline character as sentence delimiters. A manual inspection of the sentences generated using this method revealed that the newline character was not a reliable delimiter. We, therefore, only used the punctuation marks as sentence delimiters. For tokenization, we used tokenizers available in the NLTK [13] for Python. From each document we computed the number of sentences and average sentence length. The sentence length was computed as the number of words in the sentence and the average sentence length in a document is computed as the ratio of the number of words to the number of sentences. Figure 1 and
Figure **3** display the distribution of average sentence length and the number of sentences respectively for suspicious documents. Both of these figures show a positive skew indicating the presence of outliers. The outliers were traced back to the documents and a manual inspection was performed to decide if they were caused by a document processing error or if they are real values. In the suspicious documents case, the outliers were real values and the documents were kept in the corpus.

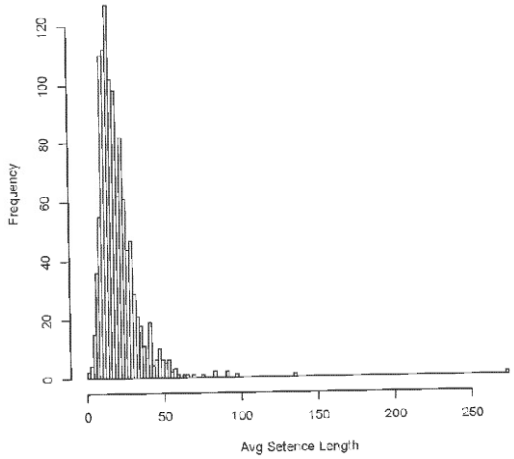


Figure 2: Distribution of average sentence length in suspicious documents

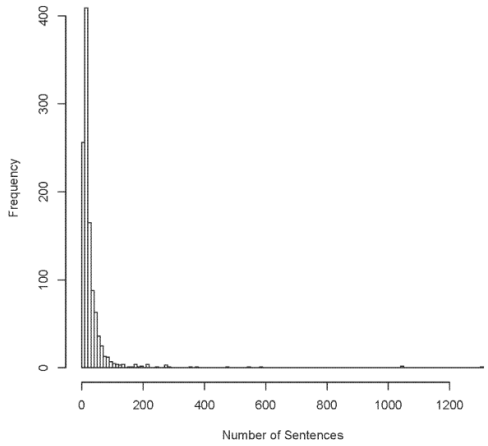


Figure 3: Distribution of number of sentences in suspicious documents

Figure **4** and
Figure **5Error! Reference source not found.** displays the same statistics for source documents. The source documents displayed similar characteristics. Unlike the suspicious documents, the outliers in the source documents were mostly caused by document processing errors such as incorrect sentence segmentation, encoding problems etc.

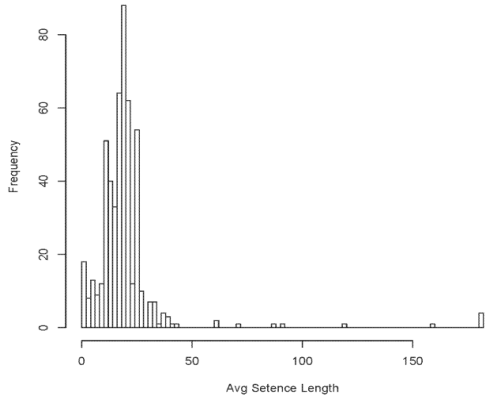


Figure 4: Distribution of average sentence length in source documents

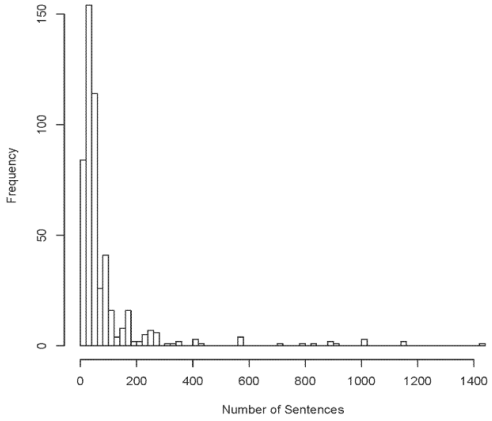


Figure 5: Distribution of number of sentences in source documents

4.1.3. Token Statistics

Apart from sentence segmentation, the documents were tokenized to collect the token level statistics from the corpus. Figure 6 and
Figure **7** display the distribution of tokens in the suspicious and source documents, respectively. Tokenization was done using the tokenizers available in NLTK.

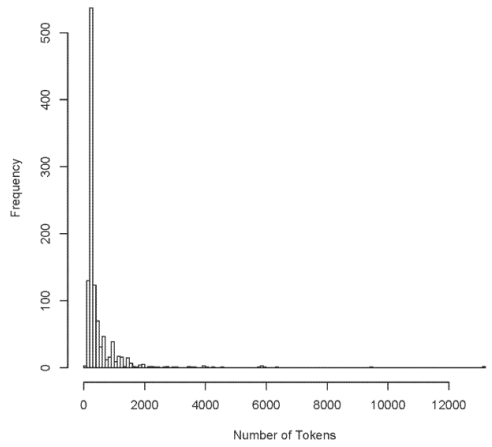


Figure 6: Distribution of the number of tokens in suspicious documents

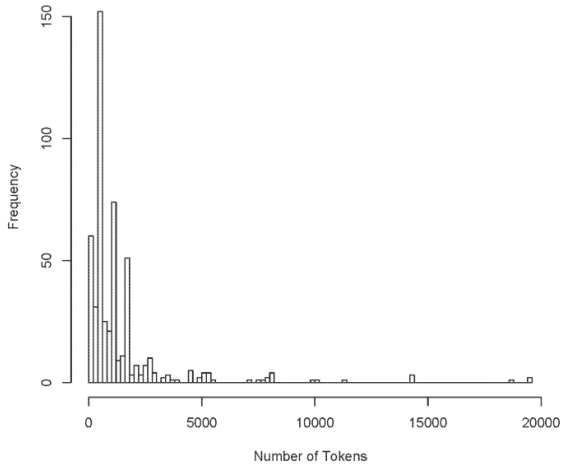


Figure 7: Distribution of the number of tokens in source documents

Table 3 displays a more detailed picture of the token statistics from the suspicious and the source documents. The source documents were, on average, much larger than the suspicious documents. This was due to the following two reasons:

1. In most of the cases, parts of the document (web page) were copied therefore the submitted assignment (suspicious document) was smaller in size compared to the web page (source document).

2. Text extraction errors as the extracted text was not limited to the main body of the web page but also included text from menus, footers and other page elements, giving the web page (source document) a larger size.

Table 3: Descriptive statistics regarding the number of tokens in the suspicious and source documents

Statistic	Suspicious	Source
Mean	519.10	1391.65
Median	282	707
Mode	201	1081
Standard Deviation	881.94	2289.77
Minimum	87	0
Maximum	13169	19572

On the other hand, the minimum size of the source document is zero indicating an error, either in the text extraction process or the unavailability of the web page altogether at the given URL. In total, we found 161 erroneous source documents, which were removed from the corpus. The final collection thus consisted of 348 suspicious document – source document pairs. The corpus also contained more than 250 documents original, non-plagiarized documents. The rest of the suspicious documents for which the source could not be obtained were removed from the final version of the corpus. The suspicious – source document pairs will be investigated for extrinsic while the non-plagiarized documents combined with plagiarized ones will be investigated for intrinsic plagiarism detection.

4. CONCLUSIONS

We developed a plagiarism detection corpus in Arabic. The corpus is annotated and organized as pairs of plagiarized – source documents along with a set of original non-plagiarized documents. Building this corpus is part of our efforts to build a plagiarism detection system for Arabic documents. We will investigate these plagiarized – source document pairs and non-plagiarized documents to investigate different intrinsic and extrinsic plagiarism detection approaches. Resources for Arabic natural language processing are fewer compared to English or other European languages. Barring any legal issues, we are planning to release the corpus for other researchers interested in investigating plagiarism in Arabic.

ACKNOWLEDGEMENTS

This work was supported by a King Abdulaziz City of Science and Technology (KACST) funding (Grant No. 11-INF-1520-03). We thank KACST for their financial support.

REFERENCES

[1] C H Leung and Y Y Chan, "A natural language processing approach to automatic plagiarism detection," in Proceedings of 8th ACME SIGITE Conference on Information Technology Education, 2007, pp. 213-218.

[2] T Wang, X Z Fan, and J Liu, "Plagiarism detection in Chinese based on chunk and paragraph weight," in Proceedings of the 7th International Conference on Machine Learning and Cybernetics, 2008, pp. 2574-2579.

[3] J A Malcolm and P C Lane, "Tackling the pan09 external plagiarism detection corpus with a desktop plagiarism detector," in Proceedings of the SEPLN, 2009, pp. 29-33.

[4] M Eissen, B Stein, and M Kulig, "Plagiarism detection without reference collections," in Proceedings of the Advances in Data Analysis, 2007, pp. 359-366.

[5] S Benno, K Moshe, and S Efstathios, "Plagiarism analysis, authorship identification and near-duplicate detection," in Proceedings of the ACM SIGIR Forum PAN07, 2007, pp. 68-71.
[6] P Clough and M Stevenson, "Developing a corpus of plagiarized short answers," Journal of Language Resources and Evaluation, vol. 45, no. 1, pp. 5-24, 2011.
[7] M Potthast, A Barrón-Cedeño, A Eiselt, B Stein, and P Rosso, "Overview of the 2nd international competition on plagiarism detection," in Notebook Papers of CLEF 2010 LABs and Workshops, 2011, pp. 19-22.
[8] D Guthrie, L Guthrie, B Allison, and Y Wilks, "Unsupervised anomaly detection," in Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007.
[9] P Clough, R Gaizauskas, S S Piao, and Y Wilks, "METER: MEasuring TExt Reuse," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 152-159.
[10] R Barzilay and K R McKeown, "Extracting paraphrases from a parallel corpus," in Proceedings of the 39 Annual Meeting of the Association of Computational Linguistics, 2001, pp. 50-57.
[11] T Cohn, C Callison-Burch, and M Lapata, "Constructing corpora for the development and evaluation of paraphrase systems," Computational Linguistics, vol. 34, no. 4, pp. 597-614, 2008.
[12] A Al-Subaihin, H Al-Khalifa, and A Al-Salman, "Sentence Boundary Detection in Colloquial Arabic Text: A Preliminary Result," in 2011 International Conference on Asian Language Processing, 2011, pp. 30-32.
[13] S Bird, E Klein, and E Loper, Natural Language Processing with Python.: O'Reilly Media, 2009.

INTENTIONAL BLANK

FRAMEWORK FOR DEVELOPED SIMPLE ARCHITECTURE ENTERPRISE – FDSAE

Nieto Bernal Wilson[1] and Luna Amaya Carmenza[2]

[1]Department of Systems Engineering, Universidad Norte-
wnieto@uninorte.edu.co
[2]Department of Industrial Engineering, Universidad Norte-
cluna@uninorte.edu.co

ABSTRACT

In This article presents a framework for develop de Architecture enterprise based on the articulation of emerging paradigms for architecture development of information enterprise [1]. The first one comes from the agile methods and it is inspired on the Scrum model which aim to simplify the complex task of developing a quality software, the second the processes models whose are oriented the development of Architectures Enterprise as Zachman and TOGAF in a paradigm of the Model Driven and principles de reference de architecture de Software form the paradigms Generation (MDG), these approaches are integrated eventually leading to the formulation and presentation of an ***framework for developed simple architecture enterprise – FDSAE-*** *The goal is to present a simple, portable, understandable terms enabling, modeling and design business information architecture in any organizational environment, in addition to this, there are important aspects related to the unified Modeling Language UML 2.5 and the Business Process Modeling BPMn that become tools to obtain the products in the FDSAE Framework, This framework is an improved version of Framework MADAIKE [2] developed by the same authors.*

KEYWORDS

Attributes Quality Performance, Architecture Enterprise, Business Process, Software Process, Metrics, Model Drive Generation, Views, Viewpoints.

1. INTRODUCTION

The construction of information systems in today's complex, to meet the demanding needs of organizations especially motivated by the technological trends, new business opportunities in the global economies, the development of business strategies, the integration of services, continuous improvement of processes in the organization and continued competition in markets increasingly require the support of computer systems to ensure functionality, security, scalability, elasticity, maintenance, capacity and availability among others. Given these new challenges software architects must have tools, processes, techniques and frameworks to resolving these challenges and get IT solutions with excellent quality attributes. In this paper display a simple and extended framework that allows to develop baseline architecture and target architecture of an organization in order to provide effective response to business needs is presented basically the FDSAE is a methodology that is based on the process abstraction and separation of layers and using modeling

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 271–284, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41222

tools can draw a scale model of the software system, on which they can make decisions and achieve launch a plan to mitigate the gaps between the current system and the ideal system within transition of the organizational. In Figure 1 the general structure shown FDSEA is proposed framework. In this one composed of seven views structure is shown: Planning, Business, Business Process, Data, Applications, Infrastructure, Safety. Accompanied by two phases of coverage, Government Enterprise Architecture and Quality Attributes (performance).

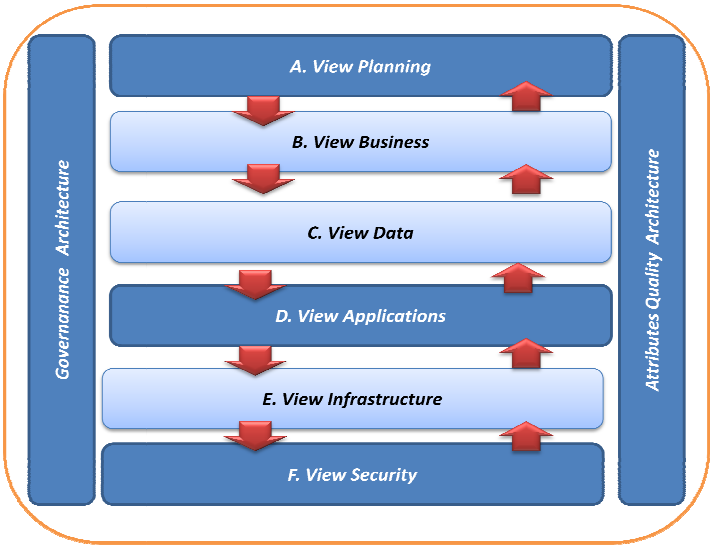


Figure 1. Methodology Structure –FDSEA-

2. RESEARCH METHODOLOGY

The following are the phases of exploratory research that conceptually possible to establish the methodology FDSEA, two phases were developed and are described as follows:

2.1. Phase I: Analysis of development standards for EA (state of the art)

This phase was aimed to evaluate the different standards, tools and process for the development of enterprise architecture (EA) on the basis of paradigm MDG (Generation Driver Modeling), in this present there are several technologies such as SysML, Zachman Framework, Togaf, SOMF, and UPDM each with different approaches and techniques to address the development of the EA. [13,14,15.16].

Model Driven Generation (MDG) Technologies: MDG Technologies allow users modeling capabilities to specific domains and notations.

Features:

- Helps align business processes and IT to the business strategies and goals.
- Provides support for at the phases in the ADM

- Provides support for OMG's Business Motivation Model · Provides support for the Architecture Content Model · Provides support for visual modeling of As-Is and To-Be architecture
- Provides support for modeling at four architecture domains specific to TOGAF (Business, Application, Data and Technology)
- Provides support for the report generation of TOGAF work products

The Systems Modeling Language (SysML): is a visual modeling language for systems engineering applications. It supports the specification, analysis, design, verification and validation of a broad range of systems. These systems may include network, infrastructure IT, hardware, software, information, processes, personnel, and facilities. SysML is a specified as a profile (dialect) of the Unified Modeling Language™ (UML™). [2],[15].

Features:

- Specify, design and analyze complex system models
- Model with all 9 diagrams for SysML.
- Visualize and trace requirements to model elements throughout the entire development lifecycle.
- Custom Search Facility: Perform complex searches, view SysML Allocations and generate reports from the results.
- Support for XMI 2.0, XMI 2.1 and UML 2.x

DDS, Data Distribution Service: DDS is a specification and an interoperability wire-protocol that defines a data-centric publishes-subscribe architecture for connecting anonymous information providers with information consumers. is a standard widely used for developing the architecture of real-time systems is the specification for a middleware type publish/subscribe distributed systems, DDS has been created with the purpose of serving the needs of industry standardize data-centric systems [2].

Features:

- Specify Data-Centric Publishers, Subscribers, Topics and QoS Policies.
- Define Data Local Reconstruction mappings for effective DDS data access.
- Target DDS implementations for the Open Splice and RTI platforms.

Framework Zachman™: The Zachman Framework™ typically is depicted as a bounded 6 x 6 “matrix” with the Communication Interrogatives as Columns and the Reification Transformations as Rows. This matrix would necessarily constitute the total set of descriptive representations that are relevant for describing something. Anything: in particular an enterprise. (www.zachman.com, 2014), The Zachman Framework™ is an Enterprise Architecture framework for enterprise architecture, which provides a formal and highly structured way of viewing and defining an enterprise. It consists of a two dimensional classification matrix based on the intersection of six communication questions (What, Where, When, Why, Who and How) with six levels of reification, successively transforming the abstract ideas on the Scope level into concrete instantiations of those ideas at the Operations level.

Features

- Displays the EA from different perspectives: Executive Management, Process, Architecture, Engineering and Technology.
- Identification, Definition, Representation, Specification, Configuration and Instantiation (Inventory, Process, Networks, Responsibility, Timing cycles and Motivation intentions.

The Open Group Architecture Framework (TOGAF): TOGAF is an architecture framework widely accepted in the industry that provides the methods and tools to assist in the acceptance, production, use and maintenance of an enterprise architecture. It is based on an iterative process model supported by best practices, and a set of reusable assets from existing architecture.

Features

- Implement all phases of the TOGAF Architecture Development Method (ADM).
- Create visual models of As-Is and To-Be architecture.
- Model all four TOGAF architecture domains: Business, Application, Data and Technology

UPDM, the Unified Profile for DoDAF/MODAF (UPDM) is the product of an Object Management Group (OMG) initiative to develop a modeling standard that supports both the USA Department of Defense Architecture Framework (DoDAF) and the UK Ministry of Defense Architecture Framework (MODAF). The MDG Technology for UPDM provides a model-based framework for planning, designing and implementing the Unified Profile for DoDAF and UPDM architectures. Source: www.sparxsystem.com, and [2].

Features

- Create architectural models for complex system-of-systems, which may include hardware, software, data, personnel and organizations.
- Define consistent, accurate architectures with clear separation of concerns to describe services, systems, operations, strategies and capabilities.
- Analyze, specify, design, and verify system models using appropriate levels of abstraction.
- Employ a rigorous, standards based approach to defining and exchanging architecture information using UML, XMI and related standards.

SOMF™ is a model-driven engineering methodology whose discipline-specific modeling language and best practices focus on software design and distinct architecture activities, employed during various stages of the software development life cycle. Moreover, architects, analysts, modelers, developers, and managers employ SOMF to tackle enterprise architecture, application architecture, service-oriented architecture (SOA), and cloud computing organizational challenges.

Features

- To achieve these underpinning milestones, six distinct software development disciplines offer corresponding models whose language notation guide practitioners in designing, architecting, and supporting a service ecosystem: Conceptual Model, Discovery and

Analysis Model, Business Integration Model, Logical Design Model, Architecture Model and Cloud Computing Toolbox Model.

2.2. Phase II: Description of components Framework for developed simple architecture enterprise –FDSAE-

2.2.1. The View Planning Architecture:

Planning view aims to develop a comprehensive project plan to address the development of the baseline architecture and target architecture of the organization, which implies:

0. Identify and understand the context of the system (organization) and defining enterprise

- Understand the proposed framework FDSAE
- Establish the scope, objectives and goals of the project.
- Establish the requirements for architecture work.
- Defining the Architecture Principles that will inform any architecture work.
- Evaluating the enterprise architecture maturity
- Estimate cost and effort.
- Identify work products to be developed
- Prepare budget and schedule
- Define and ready tools for architecture development
- Manage the project team
- Manage project risks
- Migration Plan
- Present, Socialize and obtain approval of the general project plan

Outputs: The general project plan and models preliminary of the architecture enterprise (example: Models Blocks Construction or Models of Context).

2.2.2. The View Business Architecture:

It is a layer composed of a set of models that allows visualizing the organization strategic objectives and facilitates decision making related to the integration and development of IT assets [6 and 7]. It is based on the paradigm of Enterprise Architect's Model Driven Generation (MDG) Technology, through which we obtain the following models such as: the strategic model, the Balanced Scorecard, Strategy Map, Value Chain, Decision Tree and process model organizational structures (will not be better or better organizational structure based process model organizational structure). The business architecture allows in-depth view of the system context and on this basis to establish the principles that guide the development of the following layers of the business information architecture.

Outputs: The Strategic Model, Model Chain Value, Model Organizational Structure, Model Geographic Distribution Enterprise, Model Business Process, Business Data Model and deployment General model.

Additionally other products that depend on the nature of the project and the current state of the architecture baseline may be included.

Products: Model business processes (Notation BPMn) Organizational Structure Model business (organizational structure), model high-level processes (BPMn Orchestration), Model Geographic Distribution business (model nodes), model work flows (BPMn) Model Integration work flows (BPM) with SOA, Strategy Map (Norton and Kaplan), Value Chain (Michael Porter), BSC (and Kaplan and Norton), Vision, Mission and Business Objectives.

2.2.3. View Data Architecture

A layer composed of a set of models to display the organization infrastructure integrated information, partly from conceptual models to reach the physical design of the database, data warehouses and repositories of information, Here are the models[6,7,8,]: Model of Organizational Information Objects, Logical Data Model (high level Master and Transactional), Logical Data Model (detailed), database design (physical), General Design Model and Data Warehouse Model and design of integrated repositories and. Dictionary of Data.

Outputs: Data Model Conceptual, Data Model detailed, Repository System Model, Data Warehouse and Model database design (physical),

Additionally other products, from information architecture, aims to achieve the following products: the object model Organizational Information (UML domain model) Object Model master and transactional information (object relational model), database design from information objects (physical) (database model), the Enterprise Information Integration (EEI), here is the implementation of the data model using the DBMS along with the modeling and design of Data Warehouse and Information Repositories.

2.2.4. View Applications Architecture

It is a layer composed of a set of models to display the application infrastructure, like the previous layer of conceptual models to reach physical model implementation level, in models describing this layer are: the model of high-level components, the component model and services detailed model of services, applications and software components, the integrated model (services, applications and components sw) and extended system design or distributed, (production platform).

Outputs: Component Model Conceptual, Component Model detailed, Services Model and Component design (physical),

2.2.5. Infrastructure Architecture

It is a layer composed of a set of models to visualize infrastructure hardware components, devices, connectivity networks that support communication processes, transfer of data, voice and content. Within models and systems that describe these layers are: Model business logistics system, distributed system architecture, Distributed System Design, Architecture technology and network architecture connectivity [7, 8, 9 and 10].

Outputs: Deployment Model Conceptual, Deployment Model detailed Deployment design (physical) and Style Architecture (Distributed, Virtual, Cloud and others).

Additionally other products from infrastructure architecture, aims to achieve the following products: The physical distribution model of the business (geographical node model), the model of the physical distribution business (networking) (network diagrams), the networking Design (network design), the design of the physical layout of the business (Lan, Man, Wan, Wireless, Pan) and physical layout design extended Extend Business Networking (backbone networks, transport networks, access networks and networks end user).

3. DESCRIPTION OF LAYERS, PHASES AND ROLES (FDSEA)

The FDSAE can be viewed as a multilayer model composed of six layers, see figure 1, In this case: Architecture Business, Architecture Data, Architecture functions, Architecture Infrastructure Architecture and architecture security. Plus two supporting layers made up of Governance of the architecture and Attribute Quality Architecture.

A. The View Planning Architecture: Business analysts, Board and Team architecture IT:** Mainly to assume the role of set of business analysts, Board, and team architecture IT, who is the one responsible to identify, prioritize, structure and planning architectures, defining the enterprise, Identifying key drivers and elements in the organizational context, Defining the Architecture Principles, Defining the framework to be used, Defining the relationships between management frameworks, Evaluating the enterprise architecture maturity and analyze the different topics associated with IT product benchmarking and trend analysis associated with technological product to develop.

B: Business Architecture: Business Architect IT: Mainly assumes the role of Business Architect is the one who takes care of modeling, design and business structure seen this as modeling of the structure of processes (BPM), organization, workflow, (WK) weather events (time line), identification of service components (SOA) and business strategic modeling (strategy map).

C: View Data Architecture: Information Architect: Mainly assumes the role of Information Architect is the one who is responsible for identifying and modeling Organizational Information objects (classes, objects, tables), the description of information objects both teachers and transactional (relational data model objects), design database from information objects (physical database), Enterprise Information Integration (EEI), Database, Data Repository, Data Warehouse Design and Repositories Information used in this case comprehensively the DBMS.

D: View Applications Architecture: Application Developer: Mainly it assumes the role of Application Architect is the one who is responsible for developing the architecture applications, the description of high-level software components, the modeling software components and services described the Design Component software and services detailed integration of software components and services detailed and extended system design and component-based distributed software.

E: View Infrastructure Architecture: Mainly it assumes the role of infrastructure is the architect who is responsible for modeling, design Infrastructure Architecture, the physical layout

description of the business, the modeling of the physical distribution business (networking), the design of the physical layout of the business (network), the deployment of physical distribution business (Lan, Man, Wan, Wireless, Pan) and extended physical distribution Design Business-Networking Extend-

F: View Security Architecture: Security Engineer: Mainly it assumes the role of Security Engineer, who is responsible of Identify risks architecture, define Security policies define the architecture, Organization security architecture, Identify the Security architecture users, management of the Access control architecture, Encryption and data protection architecture, Security Infrastructure, Application security and Security Compliance regulations architecture.

Parallel phase: The Governance Architecture: Business Architect: Mainly it assumes the role of Lider Team or Business Architect, who is responsible of Management a enterprise architecture consisting of business process, information, data, application and technology architecture layers for effectively and efficiently realizing enterprise and IT strategies by creating key models and practices that describe the baseline and target architectures. The architect defines and applies requirements for taxonomy, standards, guidelines, procedures, templates and tools, and provides a linkage for these components. Improve alignment, increase agility, improve quality of information and generate potential cost savings through initiatives such as re-use components the architecture.

Parallel phase: Attributes Quality Architecture: IT Architect: Mainly assumes the role of IT Architect who is the one responsible to identify, describe the Quality attributes are properties that comprehensively provide the architecture to stakeholders, are quantitative measures and qualitative the system, some examples of quality attributes by which stakeholder's assessment the quality of software systems are: Performance, security, modifiability, reliability, usability, calibrates ability, availability, throughput, configurability, Subset ability and reusability. Also the quality attributes of the software allow you to set the degree to which a software system meets its requirements for quality attributes depends on its architecture. Architectural decisions are made to promote different quality attributes. A change in architecture to promote a quality attribute often affects other quality attributes. Architecture provides the basis for achieving quality attributes, but it is useless if they did not adhere to the application.

4. DESCRIPTION OF PRODUCTS (FDSAE)

The following describes the different products to develop as a result of the application of FDSEA well: Some of the representations used in the framework of the methodology shown in figures 2, 3, 4, 5, 6, 7, 8, and 9.

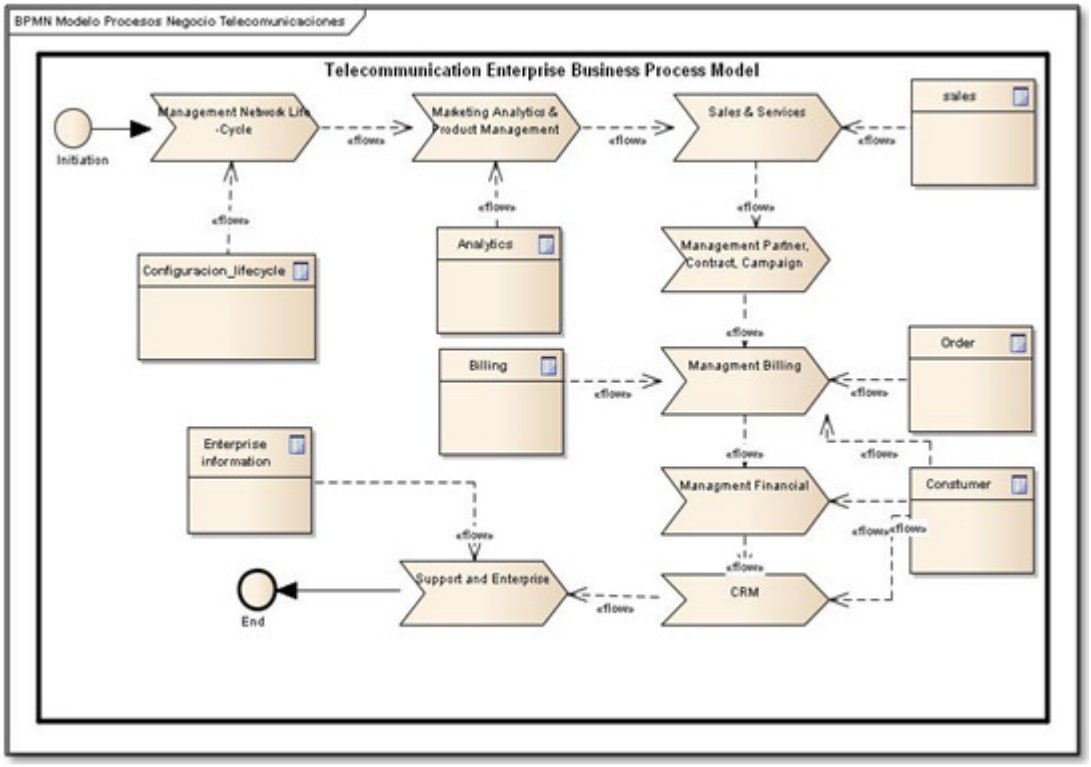


Figure 2.View Business Architecture

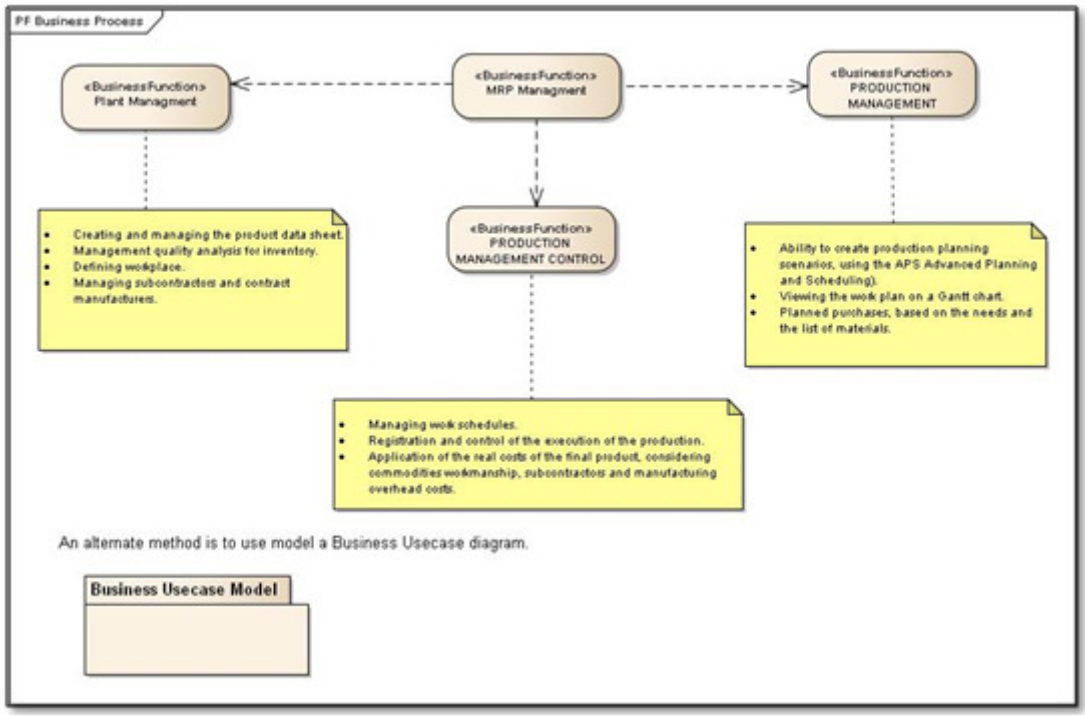


Figure 3. Plant Modeling

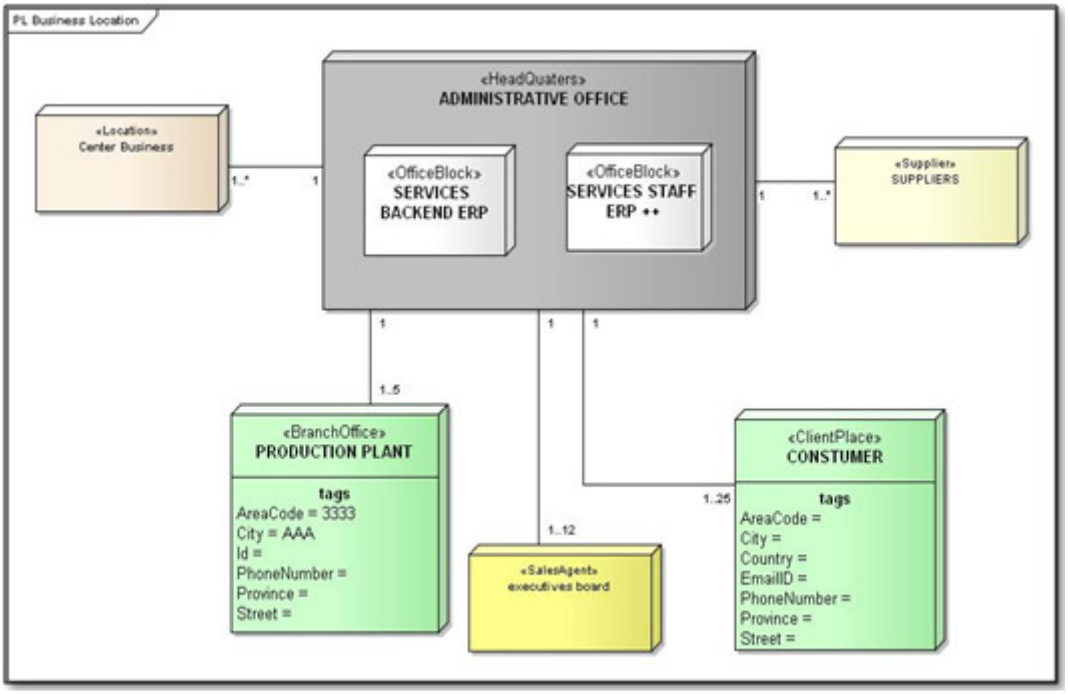


Figure 4. Geographic Distribution Business Model

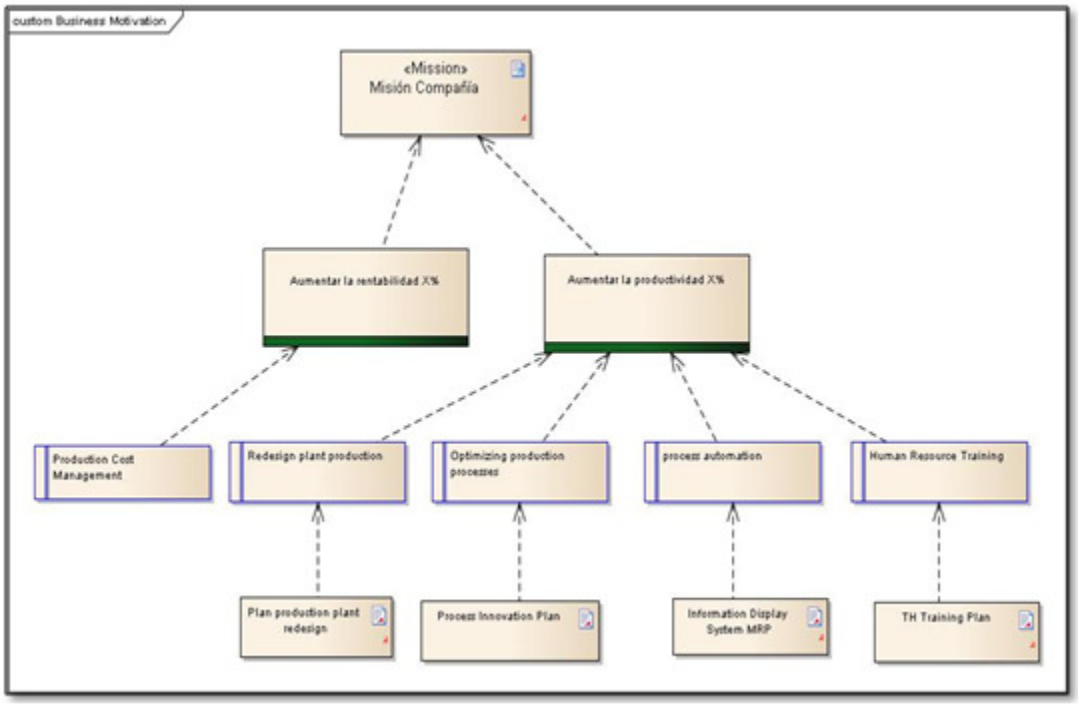


Figure 5. Motivation custom business

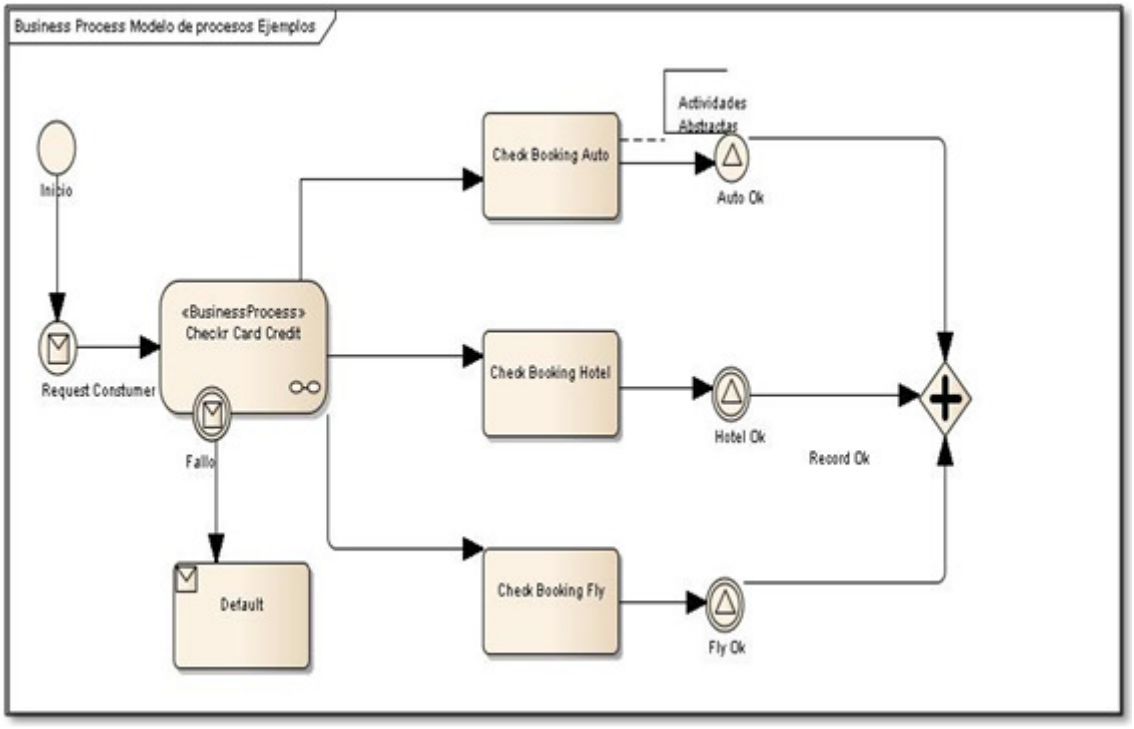


Figure 6. Activity Diagram Purchase Order Process

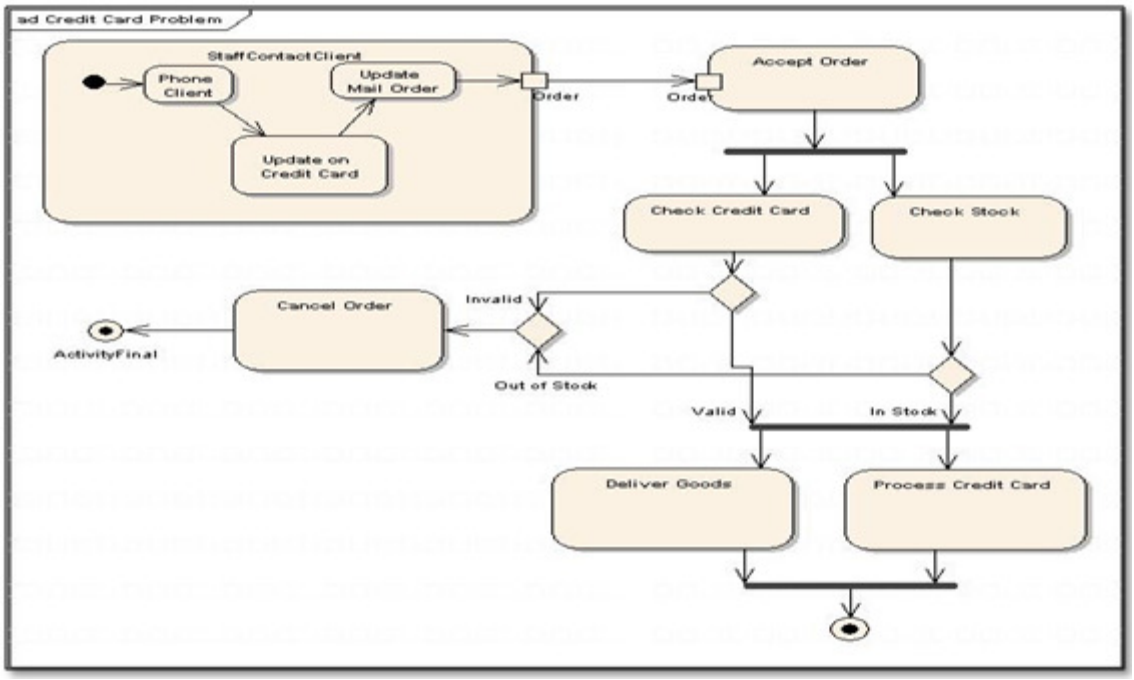


Figure 7.Business Workflow

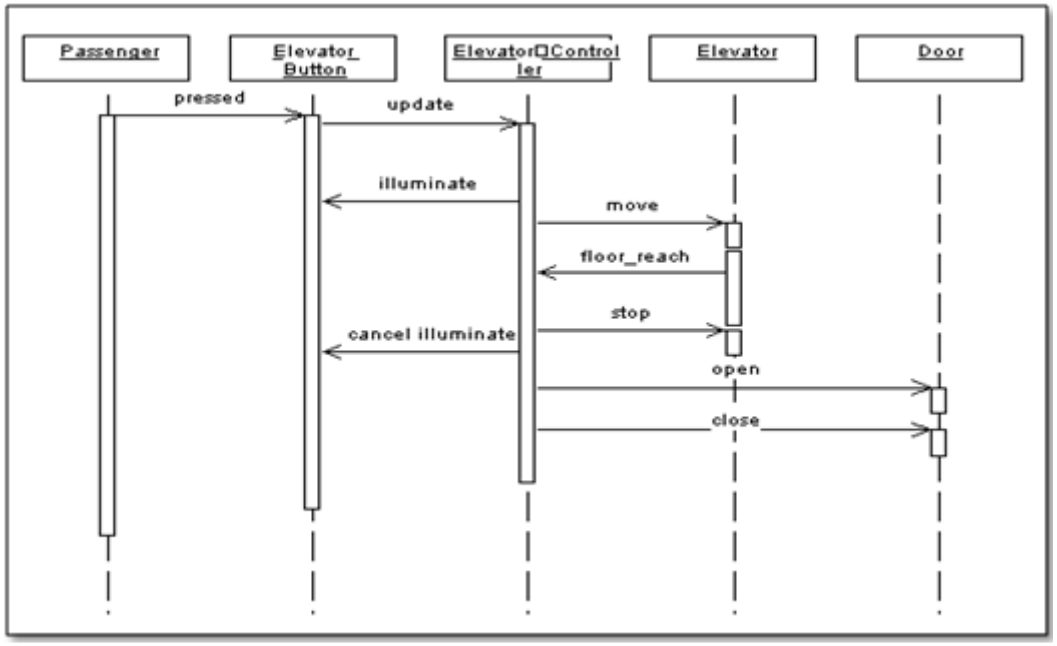


Figure 8. Sequence Diagram

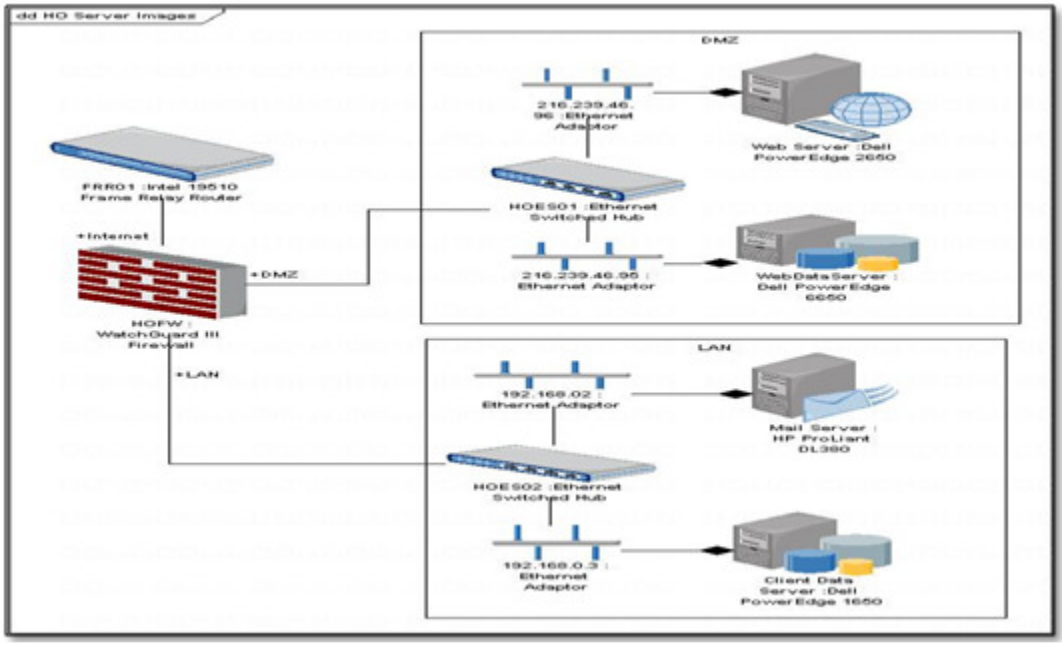


Figure 9. Deployment Model: source: www.sparxsystems.com

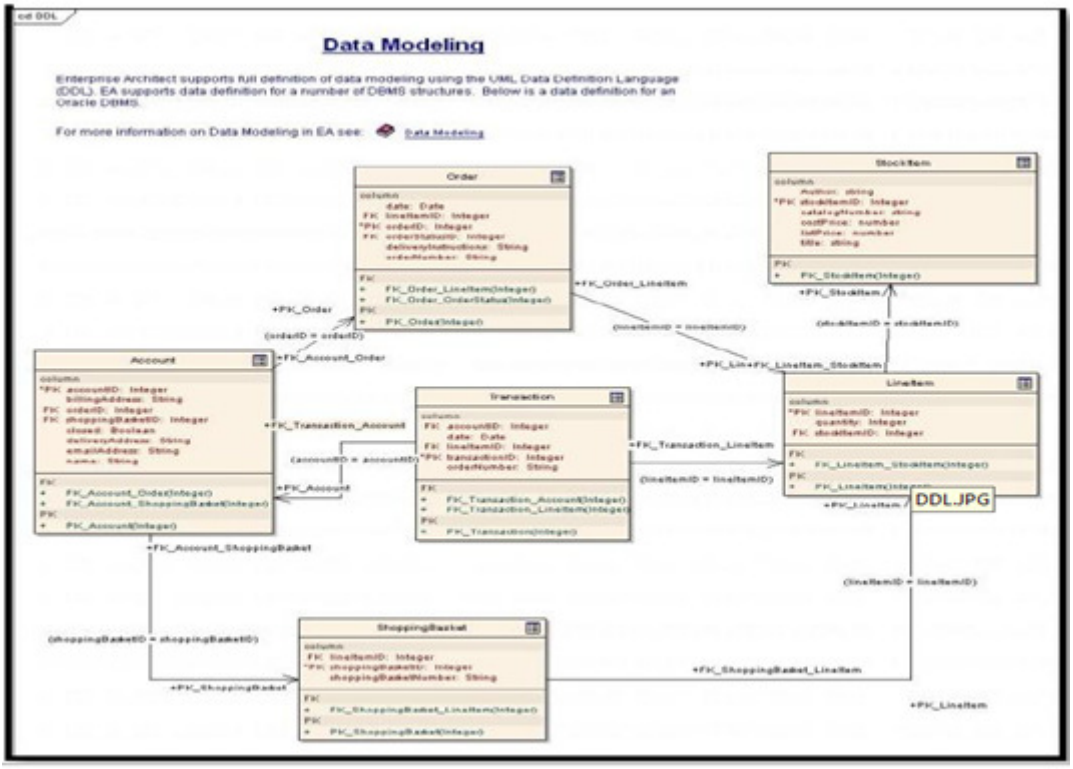


Figure 10. Domain Model source: www.sparxsystems.com

5. CONCLUSION

In this paper we demonstrate that FDSAE multilayer architecture, you can simplify complex tasks when developing quality software based on agile methods, with reference to processes oriented towards the development of enterprise information architectures which provides a standardized set of tools for project planning.

It is also important to note that the methodology objective sought to represent project phases ranging from the conceptual perspective, logic, implementation, integration and extended. In this way structure composed of layers methodology and phases resulting in a matrix model.

The methodology FDSEA can be applied to any project, regardless of size, complexity, additionally can use any tool case that supports UML 2.5 to develop different models and diagrams.

ACKNOWLEDGMENT

Financed Project with resources, “Patrimonio Autónomo Fondo Nacional de financiamiento para la Ciencia, la Tecnología y La Innovación, Francisco José De Caldas”, Proyecto 1215-502-27951 ACESCO-UNIORTE-COLCIENCAS (Colombia).

REFERENCES

[1] Godinez, M. Eberhard, H. Klaus, K. The Art of Enterprise Information Architecture, 2010, IBM Press.
[2] NietoBernal W. LunaAmaya C, Framework MADAKEI, 2014, ickse, 2014
[3] Burlton, Roger. Business Process Management: Profiting From Process. Indianapolis, IN: Sams Publishing May 2001.
[4] Chris, R. Introduction to Business Architecture, 2010, Cengage Learning.
[5] Carla Marques Pereira, Pedro Sousa, A Method to Define an Enterprise Architecture using the Zachman Framework, 2004.
[6] Crosby, Philip. Quality without Tears. New York: McGraw-Hill, 1984.
[7] Frank Goethals Jacques VandenbulckeWilfriedLemahieu, Developing the Extended Enterprise with the FADEE, 2004
[8] Gerald R. Khoury, Simeon J. Simoff, 2010, Enterprise Architecture Modelling Using Elastic Metaphors.
[9] Jeanne, W. Peter, W, Davd, C. Enterprise Architecture as Strategy, Harvard Business Press, 2006
[10] MDG Technology For Zachman, Framework User Guide, Copyright 2007-2008 Sparx Systems Pty Ltd
[11] The Open Group's approach to information systems architecture. Togaf.
[12] Model-based framework for planning, designing and implementing the Unified Profile for DoDAF and MODAF (UPDM) architectures.
[13] William Hudson, Enterprise Information Architecture: Strategies for the Real World, 2003.
[14] FDI-Intelligence, FT Business Financial Times. The fDI report 2012 Global greenfield investment trends, 2012

WEBLINKS

[15] www.sysml.org
[16] www.sparxsystem.com
[17] www.omg.org
[18] www.zachman.com
[19] www.opengroup.org/togaf/
[20] www.pmi.org

AUTHORS

Wilson Nieto Bernal is Systems Engineer and Specialist Software Engineering, Universidad Industrial de Santander (UIS) Colombia, Master /Expert in Technology Management, Master of Computer and PhD in Computer Science, ULPGC-Las Palmas GC. (Spain), 2007, with extensive experience in project management R&D in the area of Information Technology and Applications, Knowledge management Organizational Models for software development.

Carmenza Luna Amaya: is Industrial Engineering, and a PhD in Industrial Engineering from the Universidad Politecnica de Valencia, Spain, with extensive experience in research, development and organizational innovation, currently teaches at the University of the North.

DATA CHARACTERIZATION TOWARDS MODELING FREQUENT PATTERN MINING ALGORITHMS

Sayaka Akioka

School of Interdisciplinary Mathematical Sciences,
Meiji University, Tokyo, Japan
akioka@meiji.ac.jp

ABSTRACT

Big data quickly comes under the spotlight in recent years. As big data is supposed to handle extremely huge amount of data, it is quite natural that the demand for the computational environment to accelerates, and scales out big data applications increases. The important thing is, however, the behavior of big data applications is not clearly defined yet. Among big data applications, this paper specifically focuses on stream mining applications. The behavior of stream mining applications varies according to the characteristics of the input data. The parameters for data characterization are, however, not clearly defined yet, and there is no study investigating explicit relationships between the input data, and stream mining applications, either. Therefore, this paper picks up frequent pattern mining as one of the representative stream mining applications, and interprets the relationships between the characteristics of the input data, and behaviors of signature algorithms for frequent pattern mining.

KEYWORDS

Stream Mining, Frequent Mining, Characterization, Modeling, Task Graph

1. INTRODUCTION

Big data quickly comes under the spotlight in recent years. Big data is expected to collect gigantic amount of data from various data sources, and analyze those data across conventional problem domains in order to uncover new findings, or people's needs. As big data is supposed to handle extremely huge amount of data compared to the conventional applications, it is quite natural that the demand for the computational environment, which accelerates, and scales out big data applications, increases. The important thing here is, however, the behavior or characteristics of big data applications are not clearly defined yet.

Big data applications can be classified into several categories depending on the characteristics of the applications, such as behaviors, or requirements. Among those big data applications, this paper specifically focuses on stream mining applications. A stream mining application is such an application that analyzes data, which arrive one after another in chronological order, on the fly. Algorithms specialized for stream mining applications are intensively studied [1-30], and Gaber et al. published a good review paper on these algorithms [31].

Natarajan Meghanathan et al. (Eds) : WiMONe, NCS, SPM, CSEIT - 2014
pp. 285–304, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41223

High performance computing community has been investigating data intensive applications, which analyze huge amount of data as well. Raicu et al. pointed out that data intensive applications, and stream mining applications are fundamentally different from the viewpoint of data access patterns, and therefore the strategies for speed-up of data intensive applications, and stream mining applications have to be radically different [32]. Many data intensive applications often reuse input data, and the primary strategy of the speed-up is locating the data close to the target CPUs. Stream mining applications, however, rarely reuse input data, so this strategy for data intensive applications does not work in many cases. Modern computational environment has been and is evolving mainly for speed-up of benchmarks such as Linpack [33], or SPEC [34]. These benchmarks are relatively scalable according to the number of CPUs. Stream mining applications are not scalable to the contrary, and the current computational environment is not necessarily ideal for stream mining applications. The simplest approach is to use this template and insert headings and text into it as appropriate. Additionally, many researchers from machine learning domain, or data mining domain point out that the behavior, execution time more specifically, of stream mining applications varies according to the characteristics, or features of the input data. The problem is, however, the parameters, or the methodology for data characterization is not clearly defined yet, and there is no study investigating explicit relationships between the characteristics of the input data, and the behavior of stream mining applications, either.

Therefore, this paper picks up frequent pattern mining as one of the representative stream mining applications, and interprets the relationships between the characteristics of the input data, and behaviors of signature algorithms for frequent pattern mining. The rest of this paper is organized as follows. Section 2 describes a model of stream mining algorithms in order to share the awareness of the problem, which this paper focuses on. Then, the section also briefly introduces related work. Section 3 overviews the application that this paper picks up, and illustrates the algorithms those are typical solutions for the application. Section 4 explains the methodology of the experiments, shows the results, and gives discussions over those results. Section 5 concludes this paper.

2. RELATED WORK

2.1. A Model of Stream Mining Algorithms

A stream mining algorithm is an algorithm specialized for a data analysis over data streams on the fly. There are many variations of stream mining algorithms, however, general stream mining algorithms share a fundamental structure, and a data access pattern as shown in Figure 1 [35]. A stream mining algorithm consists of two parts, stream processing part, and query processing part. First, the stream processing module in stream processing part picks the target data unit, which is a chunk of data arrived in a limited time frame, and executes a quick analysis over the data unit. The quick analysis can be a preconditioning process such as a morphological analysis, or a word counting. Second, the stream processing module in stream processing part updates the data, which are cached in one or more sketches, with the latest results through the quick analysis. That is, the sketches keep the intermediate analysis, and the stream processing module updates the analysis incrementally as more data units are processed. Third, the analysis module in stream processing part reads the intermediate analysis from the sketches, and extracts the essence of the data in order to complete the quick analysis in the stream processing part. Finally, the query processing part receives this essence for the further analysis, and the whole process for the target data unit is completed.

Based on the model shown in Figure 1, we can conclude that the major responsibility of the stream processing part is to preprocess each data unit for the further analysis, and that the stream

processing part has the huge impact over the latency of the whole process. The stream processing part also needs to finish the preconditioning of the current data unit before the next data unit arrives. Otherwise, the next data unit will be lost as there is no storage for buffering the incoming data in a stream mining algorithm. On the other hand, the query processing part takes care of the detailed analysis such as a frequent pattern analysis, or a hot topic extraction based on the

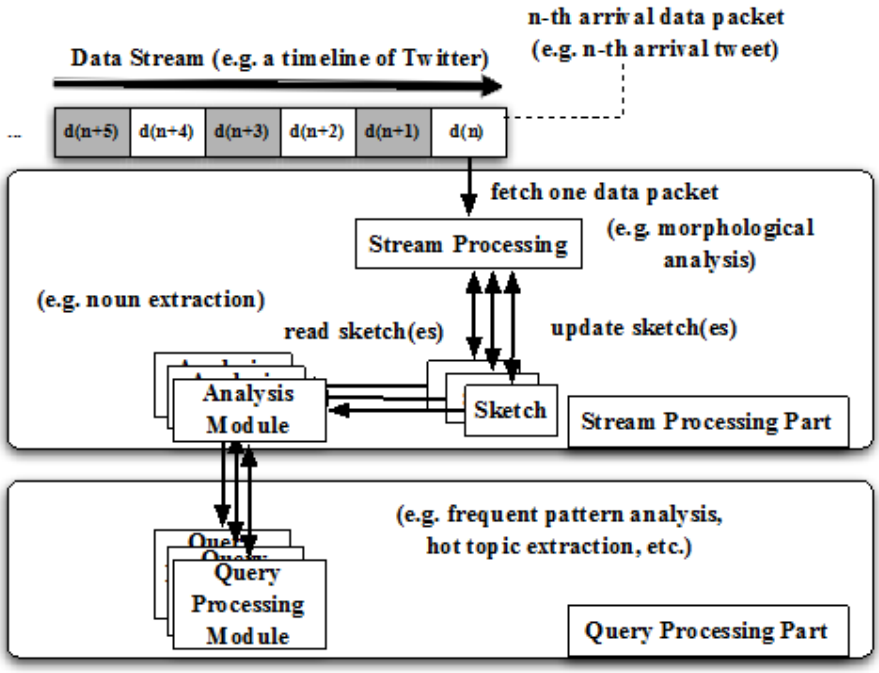


Figure 1. A model of stream mining algorithms.

intermediate data passed by the stream processing part. The output by the query processing part is usually pushed into a database system, and there is no such an urgent demand for an instantaneous response. Therefore, only the stream processing part needs to run on a real-time basis, and the successful analysis over all the incoming data simply relies on the speed of the stream processing part.

The model in Figure 1 also indicates that the data access pattern of the stream mining algorithms is totally different from the data access pattern of so-called data intensive applications, which is intensively investigated in high performance computing community. The data access pattern in the data intensive applications is a write-once-read-many [32]. That is, the application refers to the necessary data repeatedly during the computation. Therefore, the key for the speedup of the application is to place the necessary data close to the computational nodes for the faster data accesses throughout the execution of the target application. On the other hand, in a stream mining algorithm, a process refers to its data unit only once, which is a read-once-write-once style. Therefore, a scheduling algorithm for the data intensive applications is not simply applicable for the purpose of the speedup of a stream mining algorithm.

Figure 2 illustrates data dependencies between two processes analyzing data units in line, and data dependencies inside the process [35]. Here, the assumption is that each process analyzes each data unit. The left top flow represents the stream processing part of the preceding process, and the right bottom flow represents the stream processing part of the successive process. Each flow consists of six stages; read from sketches, read from input, stream processing, update sketches, read from sketches, and analysis. An arrow represents a control flow, and a dashed arrow represents a data dependency. In Figure 2, there are three data dependencies in total, and all

of these dependencies are essential to keep the analysis results consistent, and correct. The three data dependencies are as follows, and the control flow for the correct execution generates all these data dependencies.

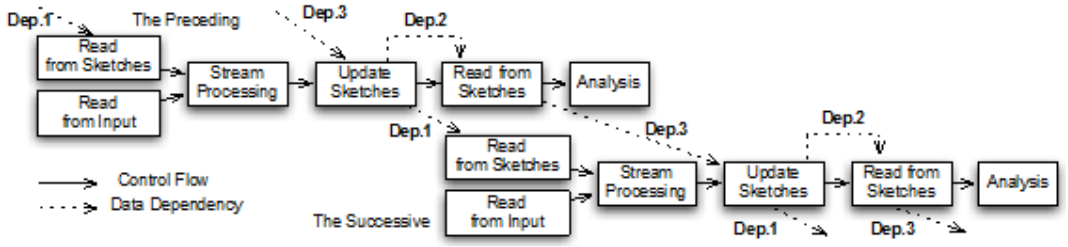


Figure 2. Data dependencies of the stream processing part in two processes in line.

- The processing module in the preceding process should finish updating the sketches before the processing module in the successive process starts reading the sketches (Dep.1 in Figure 2).

- The processing module should finish updating the sketches before the analysis module in the same process starts reading the sketches (Dep. 2 in Figure 2).

- The analysis module in the preceding process should finish reading the sketches before the processing module in the successive process starts updating the sketches (Dep. 3 in Figure 2).

2.2. A Task Graph

A task graph is a kind of pattern diagrams, which represents data dependencies, control flows, and computational costs regarding a target implementation. A task graph is quite popular for scheduling algorithm researchers. In the process of the development of scheduling algorithms, the task graph of the target implementation works as if a benchmark, and provides the way to develop a scheduling algorithm in a reproducible fashion. The actual execution of the target implementation in the actual computational environment provides realistic measurement, however, the measurement varies according to the conditions such as computational load, or timings at the moment. This fact makes difficult to compare different scheduling algorithms in order to determine which scheduling algorithm is the best for the target implementation with the target computational environment. A task graph solves this problem, and enables fair comparison of scheduling algorithms through simulations.

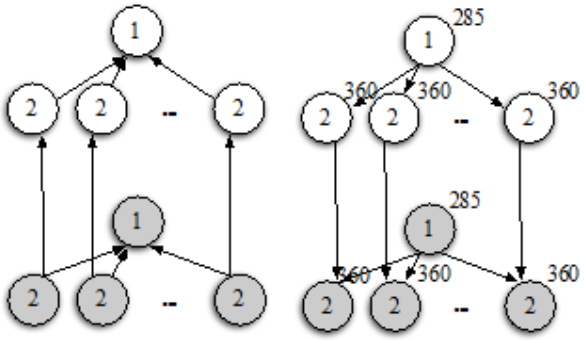


Figure 3. An example of a task graph.

for all training data **do**
  (1) fetch one data unit v
  for all attributes for v**do**
    (2-1) update the weight sum for this attribute
    (2-2) update the mean value of this attribute
  end for
end for

Figure 4 algorithm.

As discussed in the previous section, the model of a stream mining algorithm has data dependencies both across the processes, and inside one process. Therefore, a task graph for a stream mining algorithm should consist of a data dependency graph, and a control flow graph [36]. Figure 3 is an example of a task graph of the training stage of Naïve Bayes classifier [37]. In Figure 3, the left figure is a data dependency graph, and the right figure is a control flow graph. Figure 4 represents the pseudo code for the task graph shown in Figure 3.

Both a data dependency graph, and a control flow graph are directed acyclic graphs (DAGs).In a task graph, each node represents a meaningful part of the input code. Nodes in white are codes in the preceding process, and nodes in gray are codes in the successive process.The nodes with the same number represents the same meaningful part in the input source code, and the corresponding line in the pseudo code shown in Figure 4. The nodes with the same number in the same color indicate that the particular part of the input source code is runnable in parallel. Here, nodes in a control flow graph have numbers. Each number represents execution cost of the corresponding node. Each array in a data dependency graph indicates a data dependency. If an arrow comes up from node A to node B, the arrow indicates that node B relies on the data generated by node A. Similarly, each array in a control flow graph represents the order of the execution between nodes. If an arrow comes up from node A to node B, the arrow indicates that node A has to be finished before node B starts. The nodes in the same level are possible to be executed in parallel.The nodes with the same number indicate that the particular line in the pseudo code is runnable in parallel. There are two major difficulties for describing stream mining applications with task graphs. One problem is that the concrete parallelism strongly relies on the characteristics of the input data. The other problem is that a cost for a node varies according to the input data. That is, a task graph for a stream mining application is impossible without the input data modeling.

2.3. Related Work

There are several studies on task graph generation, mainly focusing on generation of random task graphs. A few projects reported task graphs generated based on the actual well-known applications; however, those applications are from numerical applications such as Fast Fourier Transformation, or other applications familiar to high performance computing community for years.

Task Graphs for Free (TGFF) provides pseudo-random task graphs [38, 39]. TGFF allows users to control several parameters, however, generates only directed acyclic graphs (DAGs) with one or multiple start nodes, and one or multiple sink nodes. A period, and deadline is assigned to each task graph based on the length of the maximum path in the graph, and the user specified parameters.

GGen is another random task graph generator proposed by Cordeiro et al. [40]. GGen generates random task graphs according to the well-known random task generation algorithms. In addition to the graph generator, GGen provides a graph analyzer, which characterizes randomly generated

task graphs, based on the longest path, the distribution of the out-degree, and the number of edges.

Task graph generator provides both random task graphs, and task graphs extracted from the actual implementations such as Fast Fourier Transformation, Gaussian Elimination, and LU Decomposition [41]. Task graph generator also provides a random task graph generator which supports a variety of network topologies including star, and ring. Task graph generator also provides scheduling algorithms as well.

Tobita et al. proposed Standard Task Graph Set (STG), evaluated several scheduling algorithms, and published the optimal schedules for STG [42, 43]. STG is basically a set of random task graphs. Tobita et al. also provided task graphs from numerical applications such as a robot control programs, a sparse matrix solver, and SPEC fpppp [34].

Besides the studies on task graph generation, Cordeiro et al. pointed out that randomly generated task graphs can create biased scheduling results, and that the biased results can mislead the analysis of scheduling algorithms [40]. According to the experiments by Cordeiro et al., a same scheduling algorithm can obtain a speedup of 3.5 times for the performance evaluation only by changing the random graph generation algorithm. Random task graphs contribute for evaluation of scheduling algorithms, however, do not perfectly cover all the domains of parallel, and distributed applications as Cordeiro et al. figured out in their work. Especially for stream mining applications, which this paper focuses on, the characteristic of the application behaviors is quite different from the characteristic of the applications familiar to the conventional high performance computing community [32]. Task graphs generated from the actual stream mining applications have profound significance in the better optimization of stream mining applications.

These all projects point out the importance of fair task graphs, and seek the best solution for this problem. These projects, however, focus only on data dependencies, control flows, and computational costs of each piece of an implementation. The computational costs are often decided in a random manner, or based on the measurements through speculative executions. No project pays attention to the relationships between the variation of the computational costs, and the characteristics of the input data.

3. ALGORITHMS

3.1. Frequent Pattern Mining

This paper focuses on two algorithms for frequent pattern mining. Frequent pattern mining was originally introduced by Agrawal et al. [44], and the baseline is the mining over the stores items purchased on a per-transaction-basis. The goal of the mining is finding out all the association rules between sets of items with some minimum specified confidence. One of the examples of the association rules is that 90% of transactions purchasing bread, and butter purchase milk as well. That is, the association rules those appearances are greater than the specified confidence are regarded as frequent patterns. Here, in the rest of this paper, the confidence is called "(minimum) support".

There are many proposals for frequent pattern mining; however, this paper specifically picks up two algorithm; Apriori algorithm [45], and FP-growth algorithm [46]. Apriori algorithm is the most basic, but standard algorithm proposed by Agrawal et al.. Many frequent mining algorithms are also developed based on Apriori algorithm. FP-growth algorithm is another algorithm, and is considered as more scalable, and faster than Apriori algorithm. The rest of this section briefly introduces summary of each algorithm.

3.2. Apriori

Figure5 gives Apriori algorithm, and there are several assumptions as follows. The items in each transaction are sorted in alphabetical order. We call the number of items in an itemset its "size", and call an itemset of size k a $k\text{-}itemset$.Items in an itemset are in alphabetical order again. We call an itemset with minimum support a "large itemset".

(1) $L_1 = large1\text{-}itemsets$
for $k \geq 2$ and $L_{k-1} \neq \emptyset$ **do**
  (2) $C_k = apriorigen(L_{k-1})$
for all transactions $t \in database$ **do**
    (3) $C_i = subset(C_k, t)$
for all candidates $c \in C_i$ **do**
      (4) c.count++
end for
end for
  (5) $L_k = \{c \in C_k | \, c.count \geq minsup\}$
end for
(6) $Answer = \cup_k L_k$

Figure 5. The pseudo-code for Apriori algorithm.

insert into C_k
select $p.item_1, p.item_2, \ldots, p.item_{k-1}, q.item_{k-1}$
from $L_{k-1}\,p, L_{k-1}\,q$
where $p.item_1 = q.item_1, \ldots, p.item_{k-2} = q.item_{k-1}, p.item_{k-1} < q.item_{k-1}$

Figure 6. The pseudo-code for the join step of apriorigen function.

for all itemsets $c \in C_k$ **do**
for all $(k-1)\text{-}subsetss$ of c **do**
if $s \notin L_{k-1}$ **then**
delete c **from** C_k
end if
end if
end if

Figure 7. The pseudo-code for the prune step of apriorigen function.

The first pass of Apriori algorithm counts item occurrences in order to determine the $large1\text{-}itemsets$ (line (1) in Figure 5). A subsequent pass consists of two phases. Suppose we are in $k\text{-}th$ pass. First, the $largeitemsetsL_{k-1}$ found in the $(k-1)\text{-}th$ pass are used for generating the candidate itemsets C_k (potentially large itemsets), using the $apriorigen$ function, which we describe later in this section (line (2) in Figure 5). Next, the database is scanned, and the support of candidates in C_k is counted (line (3) and (4) in Figure 5). These two phrases prepare for the $largeitemsetsL_k$ (line (5) in Figure 5), and the subsequent pass is repeated until L_k becomes empty.The $apriorigen$ function takes L_{k-1}, and returns a superset of the set of all the $largek\text{-}itemsets$. The $apriorigen$ function consists of the join step (Figure 6), and the prune step (Figure 7).

The distinctive part of Apriori algorithm is the $apriorigen$ function. The $apriorigen$ function reduces the size of candidate sets, and this reduction contributes for speed-up of the mining. The $apriorigen$ function is designed based on Apriori heuristic [45]; if any length k pattern is not frequent in the database, its length $k + 1$ super-pattern can never be frequent.

3.3. FP-growth

FP-growth algorithm is proposed by Han et al. [46], and they point out Apriori algorithm suffers from the cost for handling huge candidate sets, or repeated scanning database with prolific frequent patterns, long patterns, or quite low minimum support. Han et al. advocated the bottleneck exists in candidate set generation, and test, and proposed frequent pattern tree (FP-tree) as one of the alternatives.Here, we see the construction process of FP-tree, utilizing the example shown on Table 1. In this example, we set the minimum support as "three times", instead of appearance ratio in order to simplify the explanation. The FP-tree developed from this example is shown in Figure 8.

Table 1. The input example for FP-growth algorithm.

TID	item series	frequent items (for reference)
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

First, the first data scan is conducted for generating the list of frequent items. The obtained list looks like as follows.

$$< (\text{f}: 4), (\text{c}: 4), (\text{a}: 3), (\text{b}: 3), (\text{m}: 3), (\text{p}: 3) >$$

Here, $(I: c)$represents item$I$appears c times in the database. Notice the list is ordered in frequency descending order. The ordering is important as each path of a tree follows this ordering. The rightmost column on Table 1 represets the frequent items in each transaction.

Next, we start generating a tree with putting the root of a tree, labeled with "null". The second data scan is conducted for generating the FP-tree. The scan of the first transaction constructs the first branch of the tree:

$$< (\text{f}: 1), (\text{c}: 1), (\text{a}: 1), (\text{m}: 1), (\text{p}: 1) >$$

Again, notice the frequent items in the transaction are ordered in the same order to the list of frequent items. For the second transaction, its frequent item list $< f, c, a, b, m >$ shares a common prefix $< f, c, a >$ with the existing path $< f, c, a, m, p >$, we simply increment the counts of the common prefix, and create a new node $(\text{b}: 1)$ as a child of $(\text{a}: 2)$. Another new node $(\text{m}: 1)$ is created as a child of $(\text{b}: 1)$. For the third transaction, its frequent item list $< f, b >$ shares only $< f >$ with f-prefix subtree. Therefore, we increment the count of node $< f >$, and create a new node $(\text{b}: 1)$ as a child of this node $(\text{f}: 3)$. The scan for the fourth transaction

introduces a completely new branch as follows, because no node is shared with existing prefix trees.

$$< (c: 1), (b: 1), (p: 1) >$$

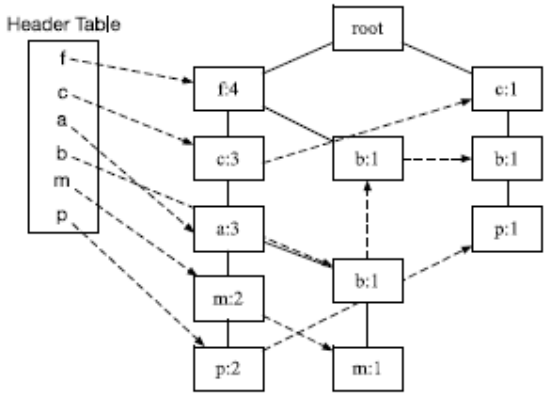


Figure 8. An example of FP-tree.Figure 4. A pseudo code for the training stage of Naïve Bayes

The frequent items in thelast transaction perfectly overlaps with the frequent items in the first transaction. Therefore, we simply increment the counts in the path$< f, c, a, m, p >$.Figure 8 is a complete FP-tree developed from this example. In the left part of Figure 8, there is an item header table, and this table contains head of node-links. This header table is utilized for tree traversal when extracting all the frequent patterns.

4. EXPERIMENTS

4.1. Setup

The purpose of this paper is to interpret the relationships between the characteristics of the input data, and behaviors of frequent pattern mining algorithms such as Apriori algorithm, and FP-growth algorithm. This section explains the methodology for the experiment for collecting data for the investigation.

We utilize the implementations of Apriori algorithm [47], and FP-growth algorithm [48] in C, which are distributed by Borgelt. We compiled, and run the programs as they are for fairness. No optimization is applied. In order to observe changes of the behaviors of these programs, three kinds of data are prepared as input. The overall feature of each data is summarized as Table 2. The details of each data are as follows.

Table 2. Summary of the input data.

	number of transactions	number of distinct items
census	48,842	135
papertitle	2,104,240	925,151
shoppers	26,496,646	836

- *census*: This is national population census data, and the data are distributed with Apriori algorithm implementation [47], and FP-growth implementation [48] by Borgelt as test data. A couple of lines from the actual data as an example are shown in Figure 9. As shown on Table 2, the ratio of the number of distinct items to the number of transactions is moderate among the three input data, and the ratio is 0.28%. We also see the number of transactions in *census* is the smallest.

- *papertitle*: This is the data set for KDD Cup 2013 Author-Paper Identification Challenge (Track 1), and available at Kaggle contest site [49]. We extract paper titles from Paper.csv, and create the list of titles for the experiments here. A couple of lines from the actual data as an example are shown in Figure 10. As shown on Table 2, the ratio of the number of distinct items to the number of transactions is extremely high compared to the other two data sets, and the ratio is 44.0%. We also see the number of transactions is moderately high.

```
1  age=middle-aged workclass=State-gov education=Bachelors edu_num=13
   marital=Never-married occupation=Adm-clerical relationship=Not-in-family
   race=White sex=Male gain=medium loss=none hours=full-time country=United-
   States salary<=50K
2  age=senior workclass=Self-emp-not-inc education=Bachelors edu_num=13
   marital=Married-civ-spouse occupation=Exec-managerial
   relationship=Husband race=White sex=Male gain=none loss=none hours=half-
   time country=United-States salary<=50K
3  age=middle-aged workclass=Private education=HS-grad edu_num=9
   marital=Divorced occupation=Handlers-cleaners relationship=Not-in-family
   race=White sex=Male gain=none loss=none hours=full-time country=United-
   States salary<=50K
```

Figure 9. Examples of census.

```
1  Stitching videos streamed by mobile phones in real-time
2  A nonlocal convection-diffusion equation
3  Area Effects in Cepaea
4  Multiple paternity in a natural population of a salamander with long-
   term sperm storage
```

Figure 10. Examples of papertitle.

```
1  707 6319 9753 2509 5555 9753 9909 5907 921 7344 4107 2106 814 9122
   4120 6315 907 9753 4509 2630 815 8101 5615 5824 907 9753 836 1908 904
   6401 3204 5620 3009 9753 3009 2301 3202 5620 2928 1905 3101 5823 3309
   1905 2908 3630 3626 3612 3319 3630 3410 3611
2  809 7113 6010 8101 2702 8101 6408
3  3601 5620 4106 6316 3307 2301 4402 2633 902 5613 811 908
```

Figure 11. Examples of shoppers.

- *shoppers*: This is the data set for Acquire Valued Shoppers Challenge, and available at Kaggle contest site as well [50]. For the experiment, we extract categories of purchased items from each transaction data in transactions.csv, and create the list of purchased item category. A couple of lines from the actual data as an example are shown in Figure 11. As shown on Table 2, the ratio of the number of distinct items to the number of transactions is extremely small compared to the other two data sets, and the ratio is 0.0032%.We also see the number of transactions is huge, and the biggest.

From the viewpoint of the application, the meaning of frequent mining over *census* is finding out typical portraits of the nation. Similarly, frequent mining over *papertitle* derives popular sets of words (not phrases) in paper titles, and frequent mining over *shoppers* extracts frequent combinations of item categories in the purchased items.

4.2. Results

Table 3 shows the number of found sets for *census* when support increases from 1.25 to 40. Both Apriori algorithm, and FP-growth algorithm found the same number of sets for the same support. Figure 12 compares the execution times in seconds of Apriori algorithm, and FP-growth algorithm for each support. Figure 13 shows the details of the execution times of Apriori algorithm for each support. Similarly, Figure 14 shows the details of the execution times of FP-growth algorithm for each support. Figure 15 shows the frequency graph of the found pattern length for each support. Here, both algorithms found the same patterns; the frequency shown in this graph is common to both Apriori algorithm, and FP-growth algorithm.

Table 3. Number of found sets (*census*).

support [%]	number of found sets
1.25	134,780
2.5	49,648
5.0	15,928
10.0	4,415
20.0	904
40.0	117

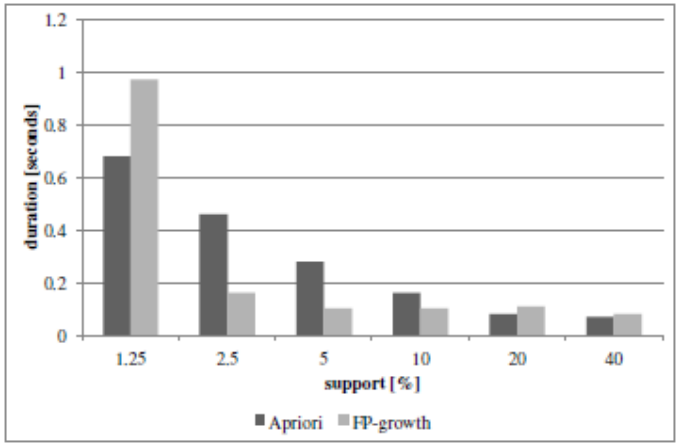


Figure 12. Comparison of durations between Apriori and FP-growth (census).

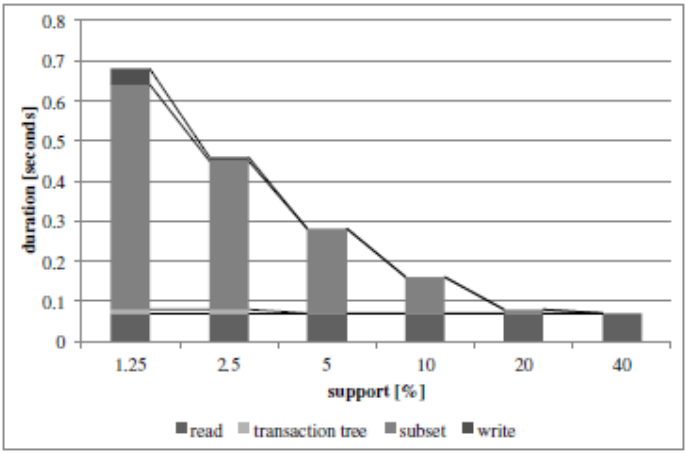


Figure 13. Details of Apriori (census).

Although the number of transactions is moderate among the three cases, the number of the found sets is quite huge compared to the other two cases (we see later on Table 4, and Table 5). This is reasonable as the ratio of the number of distinct items to the number of transactions in the input data is quite small (Table 2). According to [46], Apriori algorithm is expected to be behind because of the large number of frequent sets, and the quite low support. Figure 12 shows, however, FP-growth algorithm overcomes Apriori algorithm only when support is 2.5%, 5%, and 10%. The case with 1.25% support is supposed to be the worst case for Apriori algorithm for its low support, and the huge number of frequent sets; however, Apriori algorithm is faster than FP-growth algorithm. The details of the execution times of the two algorithms suggest the explanation of this situation. As shown in Figure 13, Apriori algorithm increases its execution time of the main part of the algorithm gently, and the curve matches to the decrease of the support. The cost for the writing part is almost constant regardless of the support contrary. Figure 14 shows FP-growth algorithm suffers from sudden and drastic increase in sorting-reducing transactions part, and writing part, and this increase caused the behind. As both Apriori algorithm, and FP-growth algorithm output the same number of sets, the sudden increase in the writing part of FP-growth algorithm apparently implicates the overhead exists in the implementation for this case. According to Han et al., Apriori algorithm suffers from the longer frequent pattern [46]. Figure 15 shows that the length of the found set is mostly centering around 12, however, scattered from 2 to 10 in the case with support value of 1.25. This tendency also indicates that the lower support value allowed the shorter pattern more, and this situation may helped Apriori algorithm, and contributed to the better performance of Apriori algorithm even with the smaller support value.

Table 4 shows the number of found sets for *papertitle* when support increases from 1.25 to 40. Both Apriori algorithm, and FP-growth algorithm found the same number of sets for the same support again. Figure 16 compares the execution times in seconds of Apriori algorithm, and FP-growth algorithm for each support.Figure 17 shows the details of the execution times of Apriori algorithm for each support. Similarly, Figure 18 shows the details of the execution times of FP-growth algorithm for each support. Figure 19 shows the frequency graph of the found pattern length for each support. Again, both algorithms found the same patterns; the frequency shown in this graph is common to both Apriori algorithm, and FP-growth algorithm.

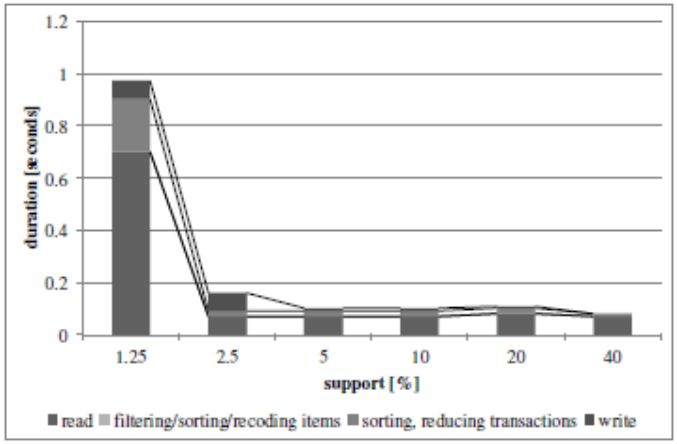


Figure 14. Details of FP-growth (census).

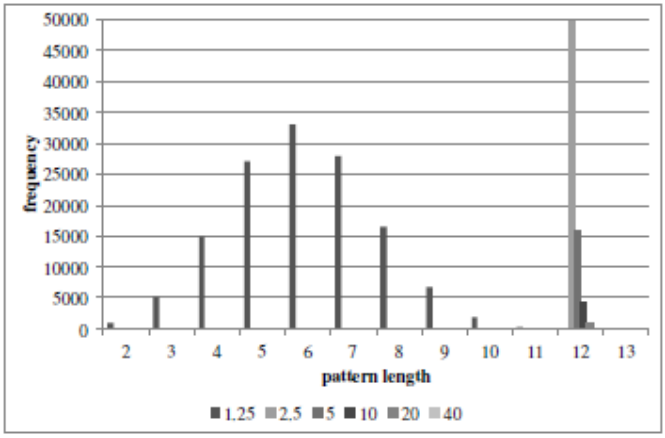


Figure 15. Pattern length (census).

Table 4. Number of found sets (*papertitle*).

support [%]	number of found sets
1.25	49
2.5	21
5.0	8
10.0	3
20.0	0
40.0	0

In this case, the input data is huge, however, the number of the found sets is relatively small as shown on Table 4. This is the reasonable output as the ratio to the number of distinct items to the number of the transactions is the highest in all the input data (Table 2).

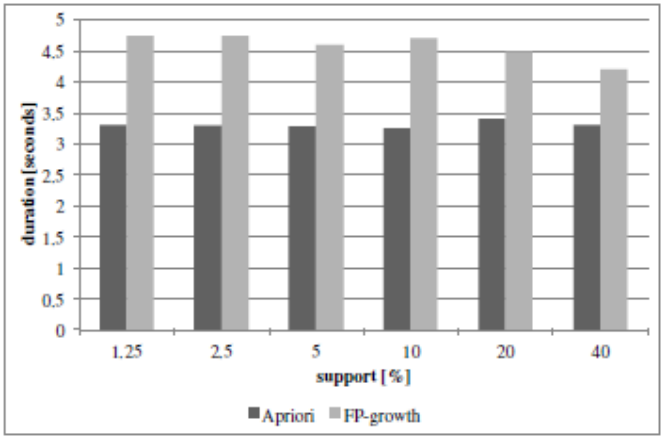


Figure 16. Comparison of durations between Apriori and FP-growth (papertitle).

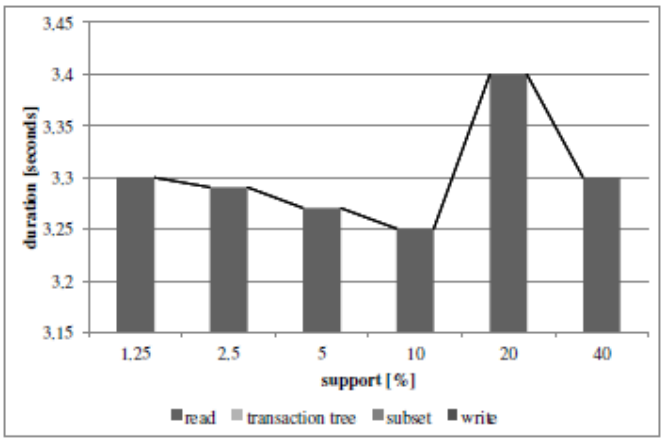


Figure 17. Details of Apriori (papertitle).

Considering the characteristics of Apriori algorithm, and FP-growth algorithm, Apriori algorithm is expected to be of advantage. Actually, Figure 16 shows Apriori algorithm is always faster than FP-growth algorithm regardless of support. Figure 17 shows Apriori algorithm does not spend meaningful time for the main part of the algorithm. Figure 18 shows FP-growth algorithm spends a certain time on the main part of the algorithm, and the impact on the overall execution time remains almost on the same level. There is no drastic increase in both writing part, and the main body of the algorithm; however, this is because the number of the found sets stays small even with the smallest support. Figure 19 shows the length of the found pattern is 2 or 3 in all the cases, which is very short, and supports the results that Apriori algorithm outperformed FP-growth algorithm.

Table 5 shows the number of found sets for *shoppers* when support increases from 1.25 to 20. For support of 40, both Apriori algorithm, and FP-growth algorithm could not list the candidate items; therefore, the experiment could not gather meaningful data. Both Apriori algorithm, and FP-growth algorithm found the same number of sets for the same support. Figure 20 compares the execution times in seconds of Apriori algorithm, and FP-growth algorithm for each support. Figure 21 shows the details of the execution times of Apriori algorithm for each support. Similarly, Figure 22 shows the details of the execution times of FP-growth algorithm for each support. Figure 23 shows the frequency graph of the found pattern length for each support. Again, both algorithms found the same patterns; the frequency shown in this graph is common to both Apriori algorithm, and FP-growth algorithm.The input data for this case is the biggest, however,

the ratio of the number of distinct items to the number of transactions is extremely small as shown on Table 5. Based on the observation of the characteristics of the input data, the number of the found sets shown in Table 5 indicates the ratio of the number of distinct items to the number of transactions is not necessarily perfect as the measure of the number of frequent sets. This is not surprising, of course, and we need to introduce some other parameters such as distribution to preconjecture the number of frequent sets from the input data. With the number of the found sets is moderately small; Apriori algorithm is always faster than FP-growth algorithm regardless of support again. Different from *papertitle*, however, both Apriori algorithm and FP-growth algorithm increase execution times as support decreases. One more thing to be considered is that Apriori algorithm increases the execution time in the main part of the algorithm as support decreases (Figure 21), however, FP-growth algorithm spends more time in reading as support decreases (Figure 22). Figure 23 also shows the length of the found sets is relatively small, which is from 2 to 4, and the situation is favorable for Apriori algorithm. This result also supports why Apriori algorithm outperformed FP-growth algorithm.

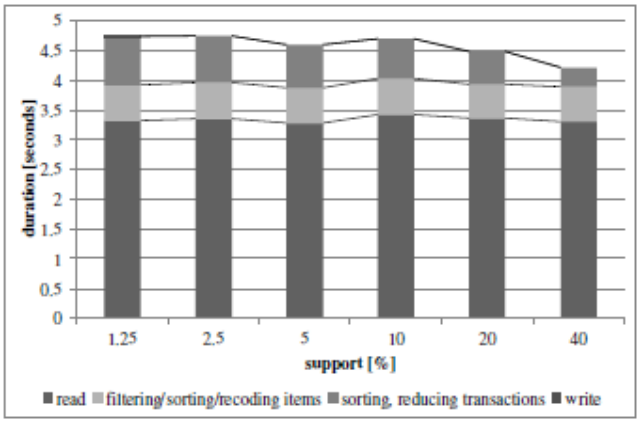


Figure 18. Details of FP-growth (papertitle).

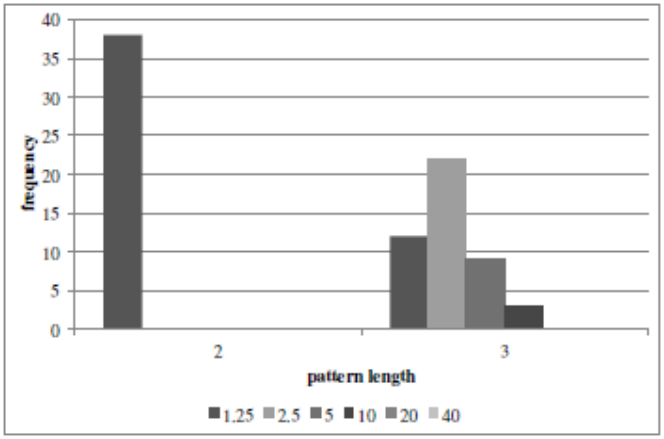


Figure 19. Pattern length (papertitle).

Table 5. Number of found sets (*shoppers*).

support [%]	number of found sets
1.25	426
2.5	69
5.0	12
10.0	1
20.0	0
40.0	-

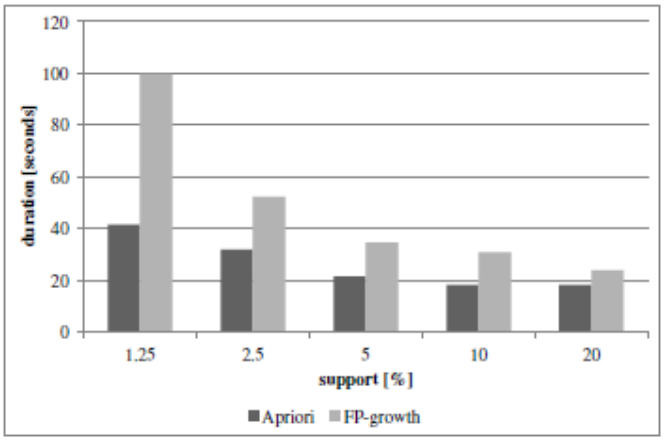


Figure 20. Comparison of durations between Apriori and FP-growth (shoppers).

4.3. Discussion

The overall results clearly indicate that simple one measure such as the total number of transactions, or the total number of distinct items does not contribute for algorithm selection. Actually, Apriori algorithm is considered to be hard to scale, however, Apriori algorithm was faster than FP-growth algorithm in *papertitle* case, and *shoppers* case, even though these two cases handle huge number of transactions compared to *census* case. Apparently, the characteristics of the input data should be determined in order to select appropriate algorithm.

As the first step of the characterization of the input data, we focused on the number of transactions, the number of distinct items, and the ratio of these two numbers. The results showed that these three parameters do contribute for the characterization; however, we need a few more parameters for the better characterization as we saw in *shoppers* case. Of course, the length of the found sets is also the important parameter. The problem is, however, those parameters such as the length of the found sets, or distributions of data, are hard to be anticipated without data mining. We need to find a good way to parameterize these characteristics.

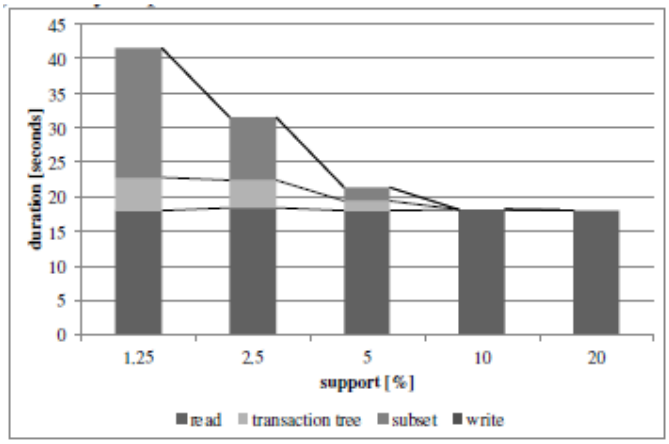


Figure 21. Details of Apriori (shoppers).

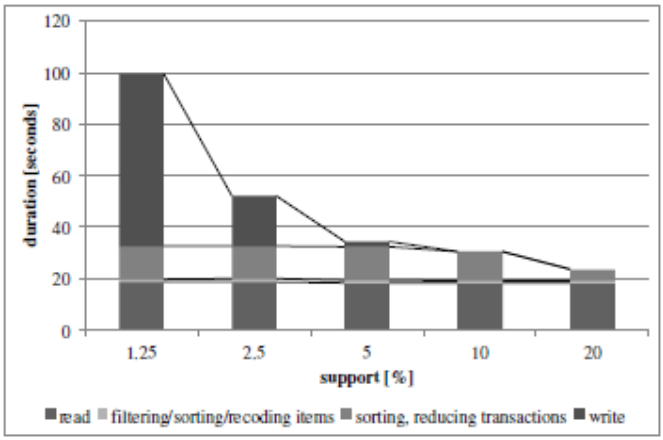


Figure 22. Details of FP-growth (shoppers).

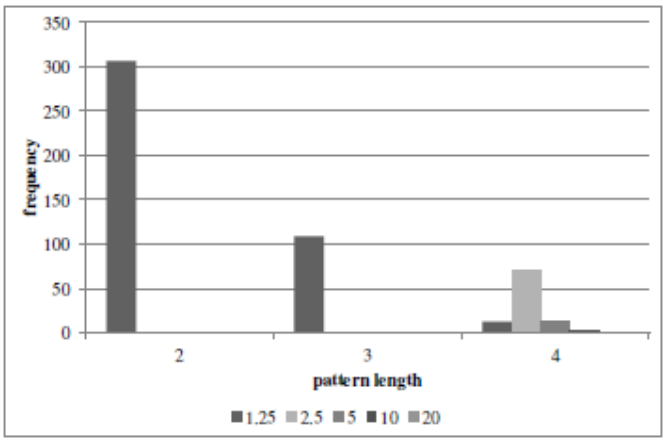


Figure 23. Pattern length (shoppers).

There is one thing left to be investigated, which is that the details of reading part and writing part of FP-growth algorithm implementation. As we observed in *census* case, and *shoppers* case, reading part, or writing part occupies large part of the overall execution time with smaller support, and the increase is not necessarily reasonable. We can make prognostications that there is something to do with reading part, or writing part is not simply doing only reading data, or

writing data, and that FP-tree has a close relation to the increase of the execution time for these two parts.

5. CONCLUSIONS

Big data quickly comes under the spotlight in recent years, and increases its importance both in socially, and scientifically. Among big data applications, this paper specifically focused on frequent mining, and gave the first step to interpret the relationships between the characteristics of the input data, and behaviors of signature algorithms for frequent pattern mining. The experiments, and discussions backed up that the characteristics of the input data have certain impact on both the performance, and the selection of the algorithm to be utilized. This paper also picked up some parameters for characterizing the input data, and showed these parameters are meaningful, however, revealed some items to be investigated for the better characterization of the input data as well.

REFERENCES

[1] B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan, "Maintaining variance and k-medians over data stream windows," in Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ser. PODS '03. New York, NY, USA: ACM, 2003, pp. 234–243.

[2] N. Tatbul, U. C¸ etintemel, S. Zdonik, M. Cherniack, and M. Stonebraker," Load shedding in a data stream manager," in Proceedings of the29th International Conference on Very Large Data Bases - Volume 29,ser. VLDB '03. VLDB Endowment, 2003, pp. 309–320.

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ser. PODS '02. New York, NY, USA: ACM, 2002, pp.1–16.

[4] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "Surfing wavelets on streams: One-pass summaries for approximate aggregate queries," in Proceedings of the 27th International Conference on VeryLarge Data Bases, ser. VLDB '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 79–88.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29thInternational Conference on Very Large Data Bases - Volume 29, ser. VLDB '03. VLDB Endowment, 2003, pp. 81–92.

[6] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for projected clustering of high dimensional datastreams," in Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30, ser. VLDB '04.VLDB Endowment, 2004, pp. 852–863.

[7] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "On demand classification of data streams," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '04. New York, NY, USA:ACM, 2004, pp. 503–508.

[8] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically," ACM Trans. Database Syst., vol. 30, no. 1, pp. 249–278, Mar. 2005.

[9] G. Dong, J. Han, L. V. S. Lakshmanan, J. Pei, H. Wang, and P. S. Yu, "Online mining of changes from data streams: Research problems and preliminary results," in Proceedings of the 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams. In cooperation with the 2003 ACM-SIGMOD International Conference on Management of Data, 2003.

[10] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan,"Clustering data streams: Theory and practice," IEEE Trans. on Knowl.and Data Eng., vol. 15, no. 3, pp. 515–528, Mar. 2003.

[11] M. Charikar, L. O'Callaghan, and R. Panigrahy, "Better streaming algorithms for clustering problems," in Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing, ser. STOC '03.New York, NY, USA: ACM, 2003, pp. 30–39. [Online].

[12] P. Domingos and G. Hulten, "Mining high-speed data streams," in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '00. NewYork, NY, USA: ACM, 2000, pp. 71–80.

[13] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing datastreams," in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '01.New York, NY, USA: ACM, 2001, pp. 97–106.

[14] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani,"Streaming-data algorithms for high-quality clustering," in Proceedings of the 18th International Conference on Data Engineering, ser. ICDE'02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 685–.

[15] C. Ordonez, "Clustering binary data streams with k-means," in Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ser. DMKD '03.New York, NY, USA: ACM, 2003, pp. 12–19.

[16] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: Implications for previous and future research," Knowl.Inf. Syst., vol. 8, no. 2, pp. 154–177, Aug. 2005.

[17] H. Wang, W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting datastreams using ensemble classifiers," in Proceedings of the Ninth ACMSIGKDD International Conference on Knowledge Discovery and DataMining, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 226–235.

[18] V. Ganti, J. Gehrke, and R. Ramakrishnan, "Mining data streams under block evolution," SIGKDD Explor. Newsl., vol. 3, no. 2, pp. 1–10, Jan.2002.

[19] S. Papadimitriou, A. Brockwell, and C. Faloutsos, "Adaptive,hands-off stream mining," in Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29, ser. VLDB'03. VLDB Endowment, 2003, pp. 560–571.

[20] M. Last, "Online classification of non stationary data streams," Intell.Data Anal., vol. 6, no. 2, pp. 129–147, Apr. 2002.

[21] Q. Ding, Q. Ding, and W. Perrizo, "Decision tree classification of spatial data streams using peano count trees," in Proceedings ofthe 2002 ACM Symposium on Applied Computing, ser. SAC '02.New York, NY, USA: ACM, 2002, pp. 413–417.

[22] G. S. Manku and R. Motwani, "Approximate frequency counts overdata streams," in Proceedings of the 28th International Conference on Very Large Data Bases, ser. VLDB '02. VLDB Endowment,2002, pp. 346–357.

[23] P. Indyk, N. Koudas, and S. Muthukrishnan, "Identifying representative trends in massive time series data sets using sketches," in Proceedings of the 26th International Conference on Very LargeData Bases, ser. VLDB '00. San Francisco, CA, USA: MorganKaufmann Publishers Inc., 2000, pp. 363–372.

[24] Y. Zhu and D. Shasha, "Statstream: Statistical monitoring of thousands of data streams in real time," in Proceedings of the 28th International Conference on Very Large Data Bases, ser. VLDB'02. VLDB Endowment, 2002, pp. 358–369.

[25] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ser. DMKD '03.New York, NY, USA: ACM, 2003, pp. 2–11.

[26] D. Turaga, O. Verscheure, U. V. Chaudhari, and L. Amini, "Resource management for networked classifiers in distributed stream mining systems," in Data Mining, 2006. ICDM '06. Sixth International Conference on, Dec 2006, pp. 1102–1107.

[27] D. S. Turaga, B. Foo, O. Verscheure, and R. Yan, "Configuring topologies of distributed semantic concept classifiers for continuous multimedia stream processing," in Proceedings of the 16th ACMInternational Conference on Multimedia, ser. MM '08. NewYork, NY, USA: ACM, 2008, pp. 289–298.

[28] B. Thuraisingham, L. Khan, C. Clifton, J. Maurer, and M. Ceruti, "Dependable real-time data mining," in Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005. Eighth IEEE International Symposium on, May 2005, pp. 158–165.

[29] N. K. Govindaraju, N. Raghuvanshi, and D. Manocha, "Fast and approximate stream mining of quantiles and frequencies using graphics processors," in Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '05. NewYork, NY, USA: ACM, 2005, pp. 611–622.

[30] K. Chen and L. Liu, "He-tree: a framework for detecting changes in clustering structure for categorical data streams," The VLDBJournal, vol. 18, no. 6, pp. 1241–1260, 2009.

[31] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining datastreams: A review," SIGMOD Rec., vol. 34, no. 2, pp. 18–26, Jun.2005.

[32] I. Raicu, I. T. Foster, Y. Zhao, P. Little, C. M. Moretti, A. Chaudhary, and D. Thain, "The quest for scalable support of data-intensive workloads in distributed systems," in Proceedings of the 18th ACM

International Symposium on High Performance Distributed Computing, ser. HPDC '09. New York, NY, USA: ACM, 2009, pp. 207–216.
[33] J. Dongarra, J. Bunch, C. Moler, and G. W. Stewart, LINPACK UsersGuide, SIAM, 1979.
[34] S. P. E. Corporation. Spec benchmarks. http://www.spec.org/benchmarks.html
[35] S. Akioka, H. Yamana, and Y. Muraoka, "Data access pattern analysis on stream mining algorithms for cloud computation," in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 2010, Las Vegas, Nevada, USA, July 12-15, 2010, 2 Volumes, 2010, pp. 36–42.
[36] S. Akioka, "Task graphs for stream mining algorithms," in Proc. The First International Workshop on Big Dynamic Distributed Data (BD32013), Riva del Garda, Italy, August 2013, pp. 55–60.
[37] K.-M. Schneider, "A comparison of event models for naive bayesanti-spam e-mail filtering," in Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics -Volume 1, ser. EACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 307–314.
[38] R. P. Dick, D. L. Rhodes, and W. Wolf, "Tgff: Task graphs for free," in Proceedings of the 6th International Workshop on Hardware/Software Codesign, ser. CODES/CASHE '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 97–101.
[39] TGFF. http://ziyang.eecs.umich.edu/_dickrp/tgff
[40] D. Cordeiro, G. Mouni´e, S. Perarnau, D. Trystram, J.-M. Vincent, and F. Wagner, "Random graph generation for scheduling simulations," in Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques, ser. SIMUTools '10. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2010, pp. 60:1–60:10.
[41] TGG, task graph generator. http://taskgraphgen.sourceforge.net
[42] STG, standard task graph set. http://www.kasahara.elec.waseda.ac.jp/schedule/index.html
[43] T. Tobita and H. Kasahara, "A standard task graph set for fair evaluation of multiprocessor scheduling algorithms," Journal of Scheduling, vol. 5, no. 5, pp. 379–394, 2002.
[44] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993ACM SIGMOD International Conference on Management of Data,ser. SIGMOD '93. New York, NY, USA: ACM, 1993, pp. 207–216.
[45] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proceedings of the 20th International Conference on Very Large Data Bases, ser. VLDB '94. San Francisco,CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.
[46] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '00.New York, NY, USA: ACM, 2000, pp. 1–12.
[47] C. Borgelt. Apriori - association rule induction/frequent item setmining. http://www.borgelt.net//apriori.html
[48] C. Borgelt. FPgrowth - frequent item set mining. http://www.borgelt.net//fpgrowth.html
[49] kaggle. Kdd cup 2013 - author-paper identification challenge (track 1). https://www.kaggle.com/c/kdd-cup-2013-author-paper-identification-challenge
[50] kaggle. Acquire valued shoppers challenge. https://www.kaggle.com/c/acquire-valued-shoppers-challenge

AUTHOR INDEX