Jasni Mohamad Zain
Wan Maseri bt Wan Mohd
Eyas El-Qawasmeh (Eds.)

Communications in Computer and Information Science 180

Software Engineering and Computer Systems

Second International Conference, ICSECS 2011
Kuantan, Pahang, Malaysia, June 2011
Proceedings, Part II

Part 2

Springer

Communications
in Computer and Information Science 180

Jasni Mohamad Zain   Wan Maseri bt Wan Mohd
Eyas El-Qawasmeh (Eds.)

Software Engineering and Computer Systems

Second International Conference, ICSECS 2011
Kuantan, Pahang, Malaysia, June 27-29, 2011
Proceedings, Part II

Springer

Volume Editors

Jasni Mohamad Zain
Wan Maseri bt Wan Mohd
Universiti Malaysia Pahang
Faculty of Computer Systems and Software Engineering
Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
E-mail: {jasni, maseri}@ump.edu.my

Eyas El-Qawasmeh
King Saud University
Information Systems Department
Riyadh 11543, Saudi Arabia
E-mail: eyasa@usa.net

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-22190-3 e-ISBN 978-3-642-22191-0
DOI 10.1007/978-3-642-22191-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930423

CR Subject Classification (1998): H.4, H.3, D.2, C.2, F.1, I.4-5

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Message from the Chairs

The Second International Conference on Software Engineering and Computer Systems (ICSECS 2011) was co-sponsored by Springer is organized and hosted by the Universiti Malaysia Pahang in Kuantan, Pahang, Malaysia, from June 27-29, 2011, in association with the Society of Digital Information and Wireless Communications. ICSECS 2011 was planned as a major event in the software engineering and computer systems field, and served as a forum for scientists and engineers to meet and present their latest research results, ideas, and papers in the diverse areas of data software engineering, computer science, and related topics in the area of digital information.

This scientific conference included guest lectures and 190 research papers that were presented in the technical session. This meeting was a great opportunity to exchange knowledge and experience for all the participants who joined us from all over the world to discuss new ideas in the areas of software requirements, development, testing, and other applications related to software engineering. We are grateful to the Universiti Malaysia Pahang in Kuantan, Malaysia, for hosting this conference. We use this occasion to express thanks to the Technical Committee and to all the external reviewers. We are grateful to Springer for co-sponsoring the event. Finally, we would like to thank all the participants and sponsors.

Jasni Mohamad Zain
Wan Maseri Wan Mohd
Hocine Cherifi

Preface

On behalf of the ICSECS 2011 Program Committee and the Universiti Malaysia Pahang in Kuantan, Pahang, Malaysia, we welcome readers to proceedings of the Second International Conference on Software Engineering and Computer Systems (ICSECS 2011).

ICSECS 2011 explored new advances in software engineering including software requirements, development, testing, computer systems, and digital information and data communication technologies. It brought together researchers from various areas of software engineering, information sciences, and data communications to address both theoretical and applied aspects of software engineering and computer systems. We do hope that the discussions and exchange of ideas will contribute to advancements in the technology in the near future.

The conference received 530 papers, out of which 205 were accepted, resulting in an acceptance rate of 39%. These accepted papers are authored by researchers from 34 countries covering many significant areas of digital information and data communications. Each paper was evaluated by a minimum of two reviewers.

We believe that the proceedings document the best research in the studied areas. We express our thanks to the Universiti Malaysia Pahang in Kuantan, Malaysia, Springer, the authors, and the organizers of the conference.

Jasni Mohamad Zain
Wan Maseri Wan Mohd
Hocine Cherifi

Organization

Program Co-chairs

Yoshiro Imai	Kagawa University, Japan
Renata Wachowiak-Smolikova	Nipissing University, Canada
Eyas El-Qawasmeh	King Saud University, Saudi Arabia

Publicity Chairs

Ezendu Ariwa	London Metropolitan University, UK
Jan Platos	VSB-Technical University of Ostrava, Czech Republic
Zuqing Zhu	University of Science and Technology of China, China

Table of Contents – Part II

Information and Data Management

Engineering

Software Security

Graphics and Multimedia

Databases

Algorithms

Signal Processings

A Mean Mutual Information Based Approach for Selecting Clustering Attribute

Hongwu Qin, Xiuqin Ma, Jasni Mohamad Zain,
Norrozila Sulaiman, and Tutut Herawan

Faculty of Computer Systems and Software Engineering,
Universiti Malaysia Pahang Lebuh Raya Tun Razak,
Gambang 26300, Kuantan, Malaysia
{qhwump,xueener}@gmail.com,
{jasni,norrozila,tutut}@ump.edu.my

Abstract. Rough set theory based attribute selection clustering approaches for categorical data have attracted much attention in recent years. However, they have some limitations in the process of selecting clustering attribute. In this paper, we analyze the limitations of three rough set based approaches: total roughness (TR), min-min roughness (MMR) and maximum dependency attribute (MDA), and propose a mean mutual information (MMI) based approach for selecting clustering attribute. It is proved that the proposed approach is able to overcome the limitations of rough set based approaches. In addition, we define the concept of mean inter-class similarity to measure the accuracy of selecting clustering attribute. The experiment results show that the accuracy of selecting clustering attribute using our method is higher than that using TR, MMR and MDA methods.

Keywords: Categorical data clustering, Mutual information, Clustering attribute.

1 Introduction

Cluster analysis is a data analysis tool used to group data with similar characteristics [1]. Recently, many researchers have contributed to categorical data clustering [1–8, 10], where data objects are made up of non-numerical attributes. Especially, rough set theory based attribute selection clustering approaches for categorical data have attracted much attention [1, 8, 10]. The key to these approaches is how to select only one attribute that is the best to cluster the objects at each time from many candidates of attributes. Mazlack *et al.* [8] proposed a technique using the average of the accuracy of approximation in the rough set theory called total roughness (TR), where the higher the total roughness is, the higher the accuracy of selecting clustering attribute. Parmar *et al.* [1] proposed a technique called Min-Min Roughness (MMR) for categorical data clustering. In selecting clustering attribute, the accuracy of approximation is measured using the well-known Marczeweski–Steinhaus metric applied to the lower and upper approximations of a subset of the universe in an information system [9]. However, MMR is the complementary of TR that produces the same accuracy

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 1–15, 2011.

and complexity with TR technique [10]. Herawan *et al.* [10] proposed a new technique called maximum dependency attributes (MDA) for selecting clustering attribute, which is based on rough set theory by taking into account the dependency of attributes of the database. Compared to TR and MMR, MDA technique provides better performance.

Due to the values of TR, MMR and MDA are all mainly determined by the cardinality of lower approximation of an attribute with respect to other attributes, they select the same attribute as clustering attribute in most of cases. Meanwhile, the three techniques have the same limitations also, that is, they tend to select the attribute with fewer values. Selecting the attribute with fewer values might cause two problems. One problem is that if an attribute has single value, it will be selected as clustering attribute, which will block the process of clustering. The other problem is that fewer clusters will be produced by such attributes, and therefore each cluster includes more objects, as a result the similarities among the objects in each cluster will be decreased.

To overcome the limitations of rough set based approaches, in this paper, we propose a mean mutual information (MMI) based approach to select clustering attribute in which the mutual information is used to measure the crispness of the clustering. In addition, we define the concept of mean inter-class similarity to measure the accuracy of selecting clustering attribute. The experiment results show that the accuracy of selecting clustering attribute using our method is higher than that using TR, MMR and MDA methods. Note that, COOLCAT [11] is also an information theory based algorithm for categorical data clustering. Different from our method, however, in [11] the entropy is used to measure the similarity among objects instead of attributes.

The rest of the paper is organized as follows. Section 2 describes rough set theory in information systems. Section 3 briefly introduces three rough set based approaches TR, MMR and MDA and analyzes the limitations of them. Section 4 describes our method MMI and gives an illustrative example. The definition of mean inter-class similarity and comparison tests of MMI with TR, MMR and MDA methods are presented in Section 5. Finally, the conclusions are described in Section 6.

2 Rough Set Theory

The notion of information system provides a convenient tool for the representation of objects in terms of their attribute values. An information system as in [12] is a 4-tuple (quadruple) $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute a, f: $U \times A \rightarrow V$ is a total function such that $f(u,a) \in V_a$, for every $(u,a) \in U \times A$, called information function. Based on the concept of information system, some basic definitions in rough set theory are given as follows.

Definition 1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ is said to be B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$.*

An indiscernibility relation induced by the set of attribute *B*, denoted by *IND* (*B*), is an equivalence relation. It is well-known that, an equivalence relation induces unique clustering. The clustering of *U* induced by *IND* (*B*) in *S* = (*U*, *A*, *V*, *f*) denoted by *U/B* and the equivalence class in the clustering *U/B* containing $x \in U$, denoted by $[x]_B$.

Definition 2. *Let S = (U, A, V, f) be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by* $\underline{B}(X)$ *and B-upper approximation of X, denoted by* $\overline{B}(X)$*, respectively, are defined by*

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \quad \textit{and} \quad \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\} \tag{1}$$

The accuracy of approximation of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted by $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{\left|\underline{B}(X)\right|}{\left|\overline{B}(X)\right|} \tag{2}$$

Throughout the paper, |*X*| denotes the cardinality of *X*.

The accuracy of approximation can also be interpreted using the well-known Marczeweski-Steinhaus (MZ) metric [9]. By applying the MZ metric to the lower and upper approximations of a subset $X \subseteq U$ in information system *S*, we have

$$D(\underline{B}(X), \overline{B}(X)) = 1 - \frac{\left|\underline{B}(X) \cap \overline{B}(X)\right|}{\left|\underline{B}(X) \cup \overline{B}(X)\right|} = 1 - \frac{\left|\underline{B}(X)\right|}{\left|\overline{B}(X)\right|} = 1 - \alpha_B(X) \tag{3}$$

Definition 3. *Let S = (U, A, V, f) be an information system and let G and H be any subsets of A. G depends on H in a degree k, is denoted by* $H \Rightarrow_k G$*. The degree k is defined by*

$$k = \frac{\sum_{X \in U/G} \left|\underline{H}(X)\right|}{|U|} \tag{4}$$

3 Analysis of Rough Set-Based Techniques

3.1 TR Technique

In an information system *S*, suppose that attribute $a_i \in A$ has *k*-different values, say β_k , *k*=1, 2, …, *n*. Let $X(a_i = \beta_k)$, *k*=1, 2, …, *n* be subsets of the objects having *k*-different values of attribute a_i. The roughness of TR technique of the set $X(a_i = \beta_k)$, k=1, 2, …, n with respect to a_j, where $i \neq j$, denoted by $R_{a_j}(X|a_i = \beta_k)$, as in [8] is defined as

$$R_{a_j}(X|a_i = \beta_k) = \frac{\left|\underline{X}_{a_j}(a_i = \beta_k)\right|}{\left|\overline{X}_{a_j}(a_i = \beta_k)\right|}, k=1, 2, \ldots, n \tag{5}$$

The mean roughness on attributes a_i with respect to a_j is defined as

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V_{a_i}|} R_{a_j}(X|a_i = \beta_k)}{|V_{a_i}|} \tag{6}$$

The total roughness of attribute $a_i \in A$ with respect to attribute a_j, where $i \neq j$, denoted by $TR(a_i)$, is obtained by the following formula

$$TR(a_i) = \frac{\sum_{j=1}^{|A|} Rough_{a_j}(a_i)}{|A| - 1} \tag{7}$$

As stated in [8] that the attribute with the highest value of TR will be selected as clustering attribute.

3.2 MMR Technique

MMR technique uses MZ metric to measure the roughness of the set $X(a_i = \beta_k)$, k=1, 2, …, n with respect to a_j, where $i \neq j$, that is

$$MMR_{a_j}(X|a_i = \beta_k) = 1 - \frac{\left|\underline{X}_{a_j}(a_i = \beta_k)\right|}{\left|\overline{X}_{a_j}(a_i = \beta_k)\right|} \tag{8}$$

The mean roughness on attributes a_i with respect to a_j is defined as

$$MMRough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V_{a_i}|} MMR_{a_j}(X|a_i = \beta_k)}{|V_{a_i}|} \tag{9}$$

The value of mean roughness is the opposite of that TR technique [10], i.e.

$$MMRough_{a_j}(a_i) = 1 - Rough_{a_j}(a_i) \tag{10}$$

Next, given *m* attributes, *Min-Roughness* (MR) of attribute $a_i (a_i \in A)$ is defined as

$$MR(a_i) = Min(MMRough_{a_1}(a_i), ..., MMRough_{a_j}(a_i), ..., MMRough_{a_m}(a_i)) \tag{11}$$

where $a_i \neq a_j$, $1 \leq j \leq m$.

MMR technique selects the attribute with the minimum value of MR as clustering attribute, i.e.

$$MMR = Min(MR(a_1), ..., MR(a_i), ..., MR(a_m)) \tag{12}$$

3.3 MDA Technique

Given any attribute *a*$_i$, *a*$_j$, $k_{a_j}(a_i)$ refer to *a*$_i$ depends on *a*$_j$ in degree *k*, is obtained by Eq. (4) as follows

$$k_{a_j}(a_i) = \frac{\sum_{X \in U/a_i} \left| \underline{a_j}(X) \right|}{|U|} \tag{13}$$

Next, *Max-Dependency* (MD) of attribute $a_i (a_i \in A)$ is defined as

$$MD(a_i) = Max(k_{a_1}(a_i), ..., k_{a_j}(a_i), ..., k_{a_m}(a_i)) \tag{14}$$

After obtaining the *m* values of $MD(a_i)$, $i = 1,2,...,m$. MDA technique selects the attribute with the maximum value of MD as clustering attribute, i.e.

$$MDA = Max(MD(a_1), ..., MD(a_i), ..., MD(a_m)) \tag{15}$$

3.4 Limitations of Rough Set Based Techniques

In [10], four test cases are used to compare and evaluate TR, MMR and MDA techniques. Four experiment results show that the three techniques always choose same attribute as the clustering attribute. This is actually not an occasional case. There is an inherent similarity among the three techniques although they look different. The similarity lies that the values of TR, MMR and MDA are all mainly determined by the cardinality of lower approximation of an attribute with respect to other attributes. There are two situations in which the cardinality of lower approximation is much higher. The first situation is that the clustering of objects induced by indiscernibility relation on one attribute is highly similar to that induced by other attributes. The other situation is that the size of domain of an attribute is relatively small. Selecting such attribute in the situation will cause two problems. To illustrate the problems, let us first see a simple example.

Example 1. Suppose we have four objects with three categorical attributes: color, shape and size as shown in Table 1.

Table 1. An objects data set

Object	Color	Shape	Size
1	red	triangle	small
2	red	circle	big
3	red	triangle	small
4	red	circle	big

Applying the TR, MMR and MDA techniques to select clustering attribute respectively, the values of TR, MMR and MDA of all attributes can be summarized in Tables 2, 3, and 4, respectively. Note that three attributes have the same MMR and MDA, so Second MMR and Second MDA are used as shown in Table 3 and Table 4.

Table 2. The total roughness of all attributes in Table 1 using TR technique

Attribute (w.r.t)	Color	Shape	Size	TR
Color	-	1	1	1
Shape	0	-	1	0.5
Size	0	1	-	0.5

Table 3. The minimum-minimum roughness of all attributes in Table 1 using MMR technique

Attribute (w.r.t)	Color	Shape	Size	MMR	Second MMR
Color	-	0	0	0	0
Shape	1	-	0	0	1
Size	1	0	-	0	1

Table 4. The degree of dependency of all attributes in Table 1 using MDA technique

Attribute (w.r.t)	Color	Shape	Size	MDA	Second MDA
Color	-	1	1	1	1
Shape	0	-	1	1	0
Size	0	1	-	1	0

According to the values of TR, MMR and MDA, the three techniques will select same attribute Color as clustering attribute. However, attribute Color has single value, selecting it as clustering attribute is obviously inappropriate because it cannot partition the objects.

Besides single-value problem, Example 1 also implies that the three techniques tend to select the attribute with fewer values. The fewer values an attribute has, the fewer clusters will be produced, and therefore each cluster includes more objects. Consequently, the similarities among the objects in each cluster will be decreased. More details about similarity and comparison tests will be described in Section 5.

4 Mean Mutual Information (MMI) Approach

To overcome the limitations of rough set based approaches, in this section we propose MMI based approach to select clustering attribute. Based on the notion of information system as stated in Section 2, we give some definitions as follows.

Definition 4. *Let S = (U, A, V, f) be an information system, if $a_i \in A$ is any attribute and $U / a_i = \{X_1, X_2, ..., X_m\}$, the entropy of a_i is defined as*

$$E(a_i) = -\sum_{i=1}^{m} P(X_i)\log_2(P(X_i)) \tag{16}$$

where $X_i \subseteq U$, $P(X_i) = \frac{|X_i|}{|U|}$.

Definition 5. *Let S = (U, A, V, f) be an information system, if $a_i, a_j \in A$ are two any attributes and $U / a_i = \{X_1, X_2, ..., X_m\}$, $U / a_j = \{Y_1, Y_2, ..., Y_n\}$, the conditional entropy (CE)of a_j with respect to a_i is defined as*

$$CE_{a_i}(a_j) = -\sum_{j=1}^{n} P(Y_j)\sum_{i=1}^{m} P(Y_j|X_i)\log_2(P(Y_j|X_i)) \tag{17}$$

where $X_i, Y_j \subseteq U$, $P(Y_j|X_i) = \frac{|Y_j \cap X_i|}{|Y_j|}$, *i=1,2,...,m and j=1,2,...,n.*

Definition 6. *Let S = (U, A, V, f) be an information system, if $a_i, a_j \in A$ are two any attributes and $U / a_i = \{X_1, X_2, ..., X_m\}$, $U / a_j = \{Y_1, Y_2, ..., Y_n\}$, the mutual information (MI)of a_j with respect to a_i is defined as*

$$MI_{a_i}(a_j) = E(a_i) - CE_{a_i}(a_j) \tag{18}$$

Definition 7. *Let S = (U, A, V, f) be an information system, if $a_i \in A$ is any attribute, the mean mutual information of a_i is defined as*

$$MMI(a_i) = \frac{\sum_{j=1, j \neq i}^{|A|} MI_{a_i}(a_j)}{|A| - 1} \tag{19}$$

In the above definitions, MI is a measurement to the similarity between two attributes a_i and a_j, concretely, the similarity between *U*/a_i and *U*/a_j. From the view of clustering, the higher MI is, the higher the crispness of the clustering. Based on these definitions, we present the MMI algorithm as shown in Fig .1.

```
Algorithm: MMI
Input: Data set
Output: Clustering attribute
Begin
  1. Compute the equivalence classes using the
     indiscernibility relation on each attribute.
  2. Compute the entropy of $a_i$ and the conditional
     entropy of all $a_j$ with respect to $a_i$, where $i \neq j$ .
  3. Compute the mutual information of all $a_j$ with
     respect to $a_i$, where $i \neq j$ .
  4. Compute the MMI of $a_i$.
  5. Select the attribute with the highest mean
     mutual information as clustering attribute.
End
```

Fig. 1. The MMI algorithm

Next, we present an illustrative example of the MMI algorithm.

Example 2. Table 5 shows an animal world data set as in [13]. There are nine animals with nine categorical attributes.

Table 5. Animal world data set from [13]

Animal	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	pointed	forward	N	claw	meat	Y	N	Y
Cheetah	Y	pointed	forward	N	claw	meat	Y	N	Y
Giraffe	Y	blunt	side	N	hoof	grass	Y	N	N
Zebra	Y	blunt	side	N	hoof	grass	Y	N	N
Ostrich	N	N	side	Y	claw	grain	N	N	N
Penguin	N	N	side	Y	web	fish	N	N	Y
Albatross	N	N	side	Y	claw	grain	N	Y	Y
Eagle	N	N	forward	Y	claw	meat	N	Y	N
Viper	N	pointed	forward	N	N	meat	N	N	N

First, we deal with attribute Hair. There are two partitions of U induced by indiscernibility relation on attribute Hair, i.e.

$$U/\text{Hair} = \{\{\text{Tiger, Cheetah, Giraffe, Zebra}\},\\ \{\text{Ostrich, Penguin, Albatross, Eagle, Viper}\}\}.$$

Applying the Eq.(16), the entropy of attribute Hair can be calculated as follows.

$$E\,(\text{Hair}) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.991$$

Applying the Eq.(17), the conditional entropy of attribute Teeth with respect to Hair can be calculated as follows.

$$CE_{\text{Hair}}\,(\text{Teeth}) = -\frac{3}{9}(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) - \frac{2}{9}\log_2 1 - \frac{4}{9}\log_2 1 = 0.306$$

Then, the mutual information of Teeth with respect to Hair is obtained by Eq.(18).

$$MI_{\text{Hair}}\,(\text{Teeth}) = E\,(\text{Hair})\text{-}\ E_{\text{Hair}}\,(\text{Teeth}) = 0.685$$

With the same process, we can get the conditional entropy and mutual information of other attributes with respect to Hair as shown in Table 6.

Table 6. CE and MI of other attributes with respect to Hair

	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
$CE_{\mathrm{Hair}}(.)$	0.306	0.984	0.401	0.539	0.444	0.000	0.766	0.984
$MI_{\mathrm{Hair}}(.)$	0.685	0.007	0.590	0.452	0.547	0.991	0.225	0.007

Subsequently, MMI of attribute Hair can be computed by Eq.(23).as

$$MMI(\mathrm{Hair}) = \frac{0.685+0.007+0.590+0.452+0.547+0.991+0.225+0.007}{8} = 0.438$$

Next, we deal with other attributes with the same process as Hair. The MI and MMI of all attributes can be summarized in Table 7.

Table 7. MI and MMI of all attributes in Table 5

Attribute	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim	MMI
Hair	-	0.685	0.007	0.590	0.452	0.547	0.991	0.225	0.007	0.438
Teeth	0.685	-	0.631	0.991	0.991	1.170	0.685	0.320	0.241	0.714
Eye	0.007	0.631	-	0.091	0.452	0.991	0.007	0.003	0.007	0.274
Feather	0.590	0.991	0.091	-	0.452	0.631	0.590	0.320	0.007	0.459
Feet	0.452	0.991	0.452	0.452	-	1.297	0.452	0.225	0.452	0.596
Eat	0.547	1.170	0.991	0.631	1.297	-	0.547	0.181	0.324	0.711
Milk	0.991	0.685	0.007	0.590	0.452	0.547	-	0.225	0.007	0.438
Fly	0.225	0.320	0.003	0.320	0.225	0.181	0.225	-	0.003	0.188
Swim	0.007	0.241	0.007	0.007	0.452	0.324	0.007	0.003	-	0.131

From Table 7, it can be seen that attribute Teeth has the highest MMI, therefore Teeth is selected as clustering attribute, which is different from the result of TR, MMR and MDA techniques.

With MMI algorithm, there is no need to worry about the problem of selecting single-value attribute as clustering attribute we mentioned in Section 2. Proposition 1 and its proof validate this view.

Proposition 1. *In an information system S = (U, A, V, f), if an attribute has single value, it is certain not to be selected as clustering attribute using MMI technique.*

Proof. Suppose attribute a_i has single value β . With MMI technique, we have,

$E(a_i) = 0$,

$CE_{a_i}(a_j) = 0$, j=1,…,|A| and $j \neq i$.

Then,

$MI_{a_i}(a_j) = 0$, j=1,…,|A| and $j \neq i$.

$MMI(a_i) = 0$.

That means a_i has the lowest value of MMI, hence it will not be selected as clustering attribute. □

5 Comparison Tests

The four techniques TR, MMR, MDA and MMI use different methods in selecting clustering attribute. Measuring the accuracy of selected clustering attribute in a just manner is a non-trivial task. In [10], total roughness stated as in Eq.(7) is used to measure the accuracy. However, TR technique selects the attribute with highest total roughness as clustering attribute; hence it is obviously unfair to other techniques using total roughness to measure the accuracy. In view of the purpose of cluster analysis is to group data with similar characteristics, we define mean inter-class similarity to measure the accuracy. Based on an information system $S = (U, A, V, f)$, we have the following definitions.

Definition 8. *Given any two objects* $x_i, x_j \in U$ *, the similarity between* x_i *and* x_j *is defined as*

$$S(x_i, x_j) = \frac{\left|\{a_k \in A \middle| f(x_i, a_k) = f(x_j, a_k)\}\right|}{|A|},\ k = 1, 2, \ldots, |A|. \tag{20}$$

Definition 9. *Suppose* $a_j \in A$ *is selected as clustering attribute and the clustering* $U / a_j = \{X_1, X_2, \ldots, X_m\}$, *where any equivalence class* $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i|X_i|}\}$. *The mean similarity (MS) of* x_{ij} *with respect to other objects is defined as*

$$MS(x_{ij}) = \frac{\sum_{k=1, k \neq j}^{|X_i|} S(x_{ij}, x_{ik})}{|X_i| - 1} \tag{21}$$

Definition 10. *Suppose* $a_j \in A$ *is selected as clustering attribute and the clustering* $U / a_j = \{X_1, X_2, \ldots, X_m\}$, *where any equivalence class* $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i|X_i|}\}$. *The inter-class similarity (CS) is defined as*

$$CS(X_i) = \frac{\sum_{j=1}^{|X_i|} MS(x_{ij})}{|X_i|} \tag{22}$$

Definition 11. *Suppose* $a_j \in A$ *is selected as clustering attribute and the clustering* $U / a_j = \{X_1, X_2, \ldots, X_m\}$, *the mean inter-class similarity (MCS) is defined as*

$$MCS(a_j) = \frac{\sum_{i=1}^{m} CS(X_i)}{m} \tag{23}$$

The higher the mean inter-class similarity is the higher the accuracy of the selected clustering attribute.

Two data sets are considered to compare and evaluate the accuracy of each technique: animal world data set from Hu [13] and the student's enrollment qualifications dataset from Herawan *et al.* [10].

5.1 Animal World Data Set in Hu [13]

Table 5 shows the animal world data set in [13]. With TR, MMR and MDA techniques, the attribute Hair is chose as clustering attribute as presented in [10]. Using MMI technique, we got the clustering attribute Teeth in Example 3. Let us measure the accuracy of attribute Hair first.

Firstly, obtain the equivalence classes induced by indiscernibility relation on attribute Hair. There are two equivalence classes as follows:

X_1 = X (Hair = Y) = { Tiger, Cheetah, Giraffe, Zebra },
X_2 = X (Hair = N) = { Ostrich, Penguin, Albatross, Eagle, Viper }.

Secondly, calculate the similarity, mean similarity, and inter-class similarity. We take animal Tiger in X_1 as an example. Applying Eq. (20), we have

S(Tiger, Cheetah) =1, *S*(Tiger, Giraffe) = 0.444, S(Tiger, Zebra) = 0.444.

Applying Eq. (21), the mean similarity of Tiger with respect to other animals in X_1 is calculated as follows:

$$MS(\text{Tiger}) = \frac{1+0.444+0.444}{3} = 0.630$$

With the same process, the similarity and mean similarity of other animals in X_1 are calculated and summarized as in Table 8.

Table 8. The similarity, MS and CS of all animals in X_1 induced by Hair

Animal	Tiger	Cheetah	Giraffe	Zebra	MS	CS
Tiger	-	1.000	0.444	0.444	0.630	0.630
Cheetah	1.000	-	0.444	0.444	0.630	
Giraffe	0.444	0.444	-	1.000	0.630	
Zebra	0.444	0.444	1.000	-	0.630	

Applying Eq. (22), the inter-class similarity of X_1 is calculated below.

$$CS(X_1) = \frac{0.630+0.630+0.630+0.630}{4} = 0.630\,.$$

Using the same way, we obtain $CS(X_2) = 0.544$.

Lastly, using Eq. (23), the mean inter-class similarity is calculated as follows:

$$MCS(\text{Hair}) = \frac{0.630 + 0.544}{2} = 0.587 .$$

Next, we measure the accuracy of Teeth obtained by MMI technique.

There are three equivalence classes induced by indiscernibility relation on attribute Teeth, i.e.

$X_1 = X$ (Teeth = Pointed) = {Tiger, Cheetah, Viper},
$X_2 = X$ (Teeth = Blunt) = {Giraffe, Zebra},
$X_3 = X$ (Teeth = N) = {Ostrich, Penguin, Albatross, Eagle}.

The similarity, mean similarity and inter-class similarity for X_1, X_2, *and X3* are calculated the same way we did on attribute Hair. The inter-class similarities of X_1, X_2, *X3* are 0.704, 1, and 0.648, respectively.

Using Eq. (23), the mean inter-class similarity is calculated as follows

$$MCS\ (\text{Teeth}) = \frac{0.704 + 1.0 + 0.648}{3} = 0.784 .$$

The result illuminates that the mean inter-class similarity of the equivalence classes induced by attribute Teeth is 0.784, which is higher than that of attribute Hair. Fig.2 illustrates the accuracy of selecting clustering attributes by four techniques.

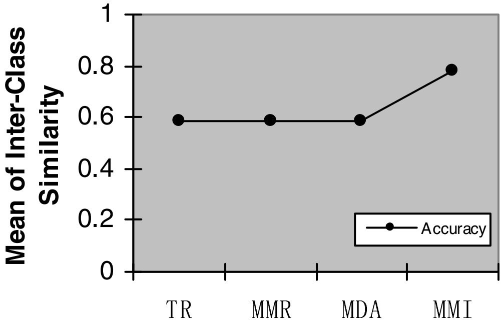


Fig. 2. The accuracy of TR, MMR, MDA and MMI techniques for animal world data set

5.2 The Student's Enrollment Qualifications in Herawan *et al.* [10]

Table 9 shows an information system of student's enrollment qualification in [10]. There are eight objects with seven categorical attributes.

Table 9. An information system of student's enrollment qualification in [10]

Student	Degree	English	Exp	IT	Math	Prog	Stat
1	Ph.D	good	medium	good	good	good	good
2	Ph.D	medium	medium	good	good	good	good
3	M.Sc	medium	medium	medium	good	good	good
4	M.Sc	medium	medium	medium	good	good	medium
5	M.Sc	medium	medium	medium	medium	medium	medium
6	M.Sc	medium	medium	medium	medium	medium	medium
7	B.Sc	medium	good	good	medium	medium	medium
8	B.Sc	bad	good	good	medium	medium	good

With TR, MMR and MDA techniques, the attribute Exp is chose as clustering attribute as presented in [10]. Using MMI technique, the mutual information and MMI of all attributes can be summarized as in Table 10.

Table 10. MI and MMI of all attributes in Table 9

Attribute	Degree	English	Exp	IT	Math	Prog	Stat	MMI
Degree	-	0.561	0.811	1.000	0.500	0.500	0.344	0.619
English	0.561	-	0.324	0.311	0.250	0.250	0.311	0.335
Exp	0.811	0.324	-	0.311	0.311	0.311	0.000	0.345
IT	1.000	0.311	0.311	-	0.000	0.000	0.189	0.302
Math	0.500	0.250	0.311	0.000	-	1.000	0.189	0.375
Prog	0.500	0.250	0.311	0.000	1.000	-	0.189	0.375
Stat	0.344	0.311	0.000	0.189	0.189	0.189	-	0.204

From Table 10, we can see that attribute Degree has the highest MMI; therefore Degree is selected as clustering attribute.

Next, let us measure the accuracy of selected clustering attributes. In this case, the process to evaluate the four techniques is the same as in animal world data set.

For attribute Exp, there are two equivalence classes:

$X_1 = X$ (Exp = medium) = {1, 2, 3, 4, 5, 6},
$X_2 = X$ (Exp = good) = {7, 8}.

Applying the Eq. (20) through Eq. (22), we obtain $CS(X_1) = 0.562$, $CS(X_2) = 0.714$. The mean inter-class similarity is calculated as follows

$$MCS\text{ (Exp)} = \frac{0.562 + 0.714}{2} = 0.638$$

For attribute Degree, there are three equivalence classes:

$X_1 = X$ (Degree = Ph.D) = {1, 2},
$X_2 = X$ (Degree = M.Sc) = {3, 4, 5, 6},
$X_3 = X$ (Degree = B.Sc) = {7, 8}.

Similarly, we obtain $CS(X_1) = 0.857$, $CS(X_2) = 0.738$, and $CS(X_1) = 0.714$, respectively. The mean inter-class similarity is calculated as follows

$$MCS\,(\text{Degree}) = \frac{0.857 + 0.738 + 0.714}{3} = 0.770$$

The result illuminates that the mean inter-class similarity of the equivalence classes induced by attribute Degree is 0.770, which is higher than that of attribute Exp. Fig.3 illustrates the accuracy of selecting clustering attributes by four techniques.

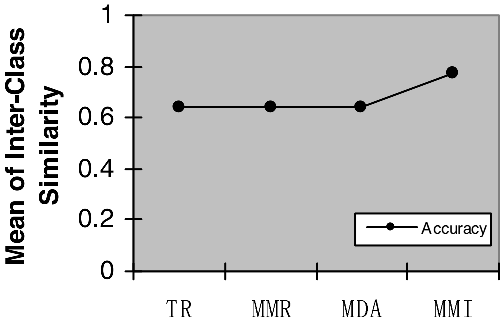


Fig. 3. The accuracy of TR, MMR, MDA and MMI techniques for student's enrollment qualification data set

The results of above two experiments show that the accuracy of selecting clustering attribute using MMI technique is higher than that using TR, MMR and MDA techniques.

6 Conclusion

Currently, applying rough set theory in the process of selecting clustering attribute is one of popular approaches. However, there are some inherent limitations in the rough set based method. In this paper, we analyze the limitations of three rough set based methods: total roughness (TR), min-min roughness (MMR) and maximum dependency attribute (MDA), and propose a mean mutual information (MMI) based method for selecting clustering attribute. It is proved that the proposed approach can overcome the limitations of rough set based method. In addition, we define the concept of mean inter-class similarity to measure the accuracy of selecting clustering attribute. The experiment results on two data sets shows that the accuracy of selecting clustering attribute using our method is higher than that using TR, MMR and MDA methods. The proposed approach could be integrated into clustering algorithm based on attributes selection for categorical data.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Parmar, D., Wu, T., Blackhurst, J.: MMR: an algorithm for clustering categorical data using rough set theory. Data and Knowledge Engineering 63, 879–893 (2007)
2. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categori-cal values. Data Mining and Knowledge Discovery 2(3), 283–304 (1998)
3. Gibson, D., Kleinberg, J., Raghavan, P.: Clustering categorical data: an approach based on dynamical systems. The Very Large Data Bases Journal 8(3-4), 222–236 (2000)
4. Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. Information Systems 25(5), 345–366 (2000)
5. Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS –clustering categorical data using summaries. In: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 73–83 (1999)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
7. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science 11, 341–356 (1982)
8. Mazlack, L.J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing clustering attributes. In: Proceedings of the ISCA 13th International Conference (CAINE 2000), pp. 1–6 (2000)
9. Yao, Y.Y.: Information granulation and rough set approximation. International Journal of Intelligent Systems 16(1), 87–104 (2001)
10. Herawan, T., Deris, M.M., Abawajy, J.H.: A rough set approach for selecting clustering attribute. Knowledge-Based Systems 23, 220–231 (2010)
11. Barbara, D., Li, Y., Couto, J.: COOLCAT: an entropy-based algorithm for categorical clustering. In: Proc. of CIKM 2002, pp. 582–589 (2002)
12. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences: An International Journal 177(1), 3–27 (2007)
13. Hu, X.: Knowledge discovery in databases: an attribute oriented rough set approach. Ph.D. Thesis, University of Regina (1995)

A Soft Set Model on Information System and Its Application in Clustering Attribute Selection

Hongwu Qin, Xiuqin Ma, Jasni Mohamad Zain,
Norrozila Sulaiman, and Tutut Herawan

Faculty of Computer Systems and Software Engineering,
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
{qhwump,xueener}@gmail.com,
{jasni,norrozila,tutut}@ump.edu.my

Abstract. In this paper, we define a soft set model on the set of equivalence classes in an information system, which can be easily applied to obtaining approximation sets of rough set. Furthermore, we use it to select clustering attribute for categorical data clustering and a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

Keywords: Soft set, Rough set, Information system, Clustering attribute.

1 Introduction

In 1999, Molodtsov [1] proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. Where soft set theory is different from traditional tools for dealing with uncertainties, such as the theory of probability, the theory of fuzzy sets, is that it is free from the inadequacy of the parametrization tools of those theories. At present, work on the soft set theory is progressing rapidly both in theoretical models and applications. Recently, the relation between the rough set and soft set has also attracted much attention. Feng et al. [4] investigated the problem of combining soft sets with fuzzy sets and rough sets. Three different types of hybrid models were presented, which were called rough soft sets, soft rough sets and soft-rough fuzzy sets, respectively. Herawan and Mat Deris give a direct proof in [5] that every rough set is a soft set.

Rough set theory [2], introduced by Z. Pawlak in 1982, is a mathematical tool to deal with vagueness and uncertainty. It has been widely used in many branches of artificial intelligence and data mining. The original goal of the rough set theory is induction of approximations of concepts. The idea consists of approximation of a subset by a pair of two precise concepts called the lower approximation and upper approximation. Intuitively, the lower approximation of a set consists of all elements that surely belong to the set, whereas the upper approximation of the set constitutes of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 16–27, 2011.

all elements that possibly belong to the set. The difference of the upper approximation and the lower approximation is a boundary region. It consists of all elements that cannot be classified uniquely to the set or its complement, by employing available knowledge.

In this paper, we propose a soft set model on information system. The soft set model is constructed over the set of equivalence class instead of the set of single object. Then we apply the soft set model to obtaining approximation sets of rough set. Furthermore, we use it to select clustering attribute for categorical data cluster and a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

The rest of this paper is organized as follows. The following section briefly reviews some basic definitions in rough set theory and soft sets. Section 3 describes the construction of soft set model on information system. Section 4 shows the application of the proposed model in Selecting Clustering Attributes. Section 5 makes comparison between MDA approach and the proposed technique. Finally, conclusions are given in Section 6.

2 Preliminaries

The notion of information system provides a convenient tool for the representation of objects in terms of their attribute values. An information system as in [3] is a 4-tuple $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \bigcup_{a\in A} V_a$, V_a is the domain (value set) of attribute a, f: $U \times A \rightarrow V$ is a total function such that $f(u,a) \in V_a$, for every $(u,a) \in U \times A$, called information function. Next, we review some basic definitions with regard to rough set and soft set.

Definition 1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ is said to be B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$.*

An indiscernibility relation induced by the set of attribute B, denoted by IND (B), is an equivalence relation. It is well-known that, an equivalence relation induces unique partition. The partition of U induced by IND (B) in $S = (U, A, V, f)$ denoted by U/B and the equivalence class in the partition U/B containing $x \in U$, denoted by $[x]_B$.

Definition 2. *Let $S = (U, A, V, f)$ be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by $\underline{B}(X)$ and B-upper approximation of X, denoted by $\overline{B}(X)$, respectively, are defined by*

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\} \tag{1}$$

The accuracy of approximation of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted by $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \tag{2}$$

Throughout the paper, |*X*| denotes the cardinality of *X*.

Definition 3. *Let S = (U, A, V, f) be an information system and let G and H be any subsets of A. G depends on H in a degree k, is denoted by* $H \Rightarrow_k G$. *The degree k is defined by*

$$k = \frac{\sum_{X \in U/G} |\underline{H}(X)|}{|U|} \tag{3}$$

Let *U* be an initial universe of objects, *E* be the set of parameters in relation to objects in *U*, *P*(*U*) denotes the power set of *U*. The definition of soft set is given as follows.

Definition 4. *A pair (F, E) is called a soft set over U, where F is a mapping given by*

$$F: E \rightarrow P(U)$$

From the definition, a soft set (*F*, *E*) over the universe *U* is a parameterized family of subsets of the universe *U*, which gives an approximate description of the objects in *U*. For any parameter $e \in E$, the subset $F(e) \subseteq U$ may be considered as the set of *e*-approximate elements in the soft set (*F*, *E*).

Example 1. Let us consider a soft set (*F*, *E*) which describes the “attractiveness of houses” that Mr. X is considering to purchase. Suppose that there are six houses in the universe $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ under consideration and E = $\{e_1, e_2, e_3, e_4, e_5\}$ is the parameter set, where e_i (i =1,2,3,4,5) stands for the parameters “beautiful”, “expensive”, “cheap”, “good location” and “wooden” respectively. Consider the mapping *F: E* → *P*(*U*) given by “houses (.)”, where (.) is to be filled in by one of parameters $e \in E$. Suppose that $F(e_1)=\{h_1, h_3, h_6\}$, $F(e_2)=\{h_1, h_2, h_3, h_6\}$, $F(e_3)=\{h_4, h_5\}$, $F(e_4)=\{h_1, h_2, h_6\}$, $F(e_5)=\{h_5\}$. Therefore, $F(e_1)$ means “houses (beautiful)”, whose value is the set $\{h_1, h_3, h_6\}$.

In order to facilitate storing and dealing with soft set, the binary tabular representation of soft set is often given in which the rows are labeled by the object names and columns are labeled by the parameter names, and the entries are denoted by $F(e_j)(x_i), (e_j \in E, x_i \in U, j = 1,2,..m, x = 1,2,..n)$. If $x_i \in F(e_j)$, then $F(e_j)(x_i)$ =1, otherwise $F(e_j)(x_i)$ =0. Table 1 is the tabular representation of the soft set (*F*, *E*) in Example 1.

Table 1. Tabular representation of the soft set (*F*, *E*)

U	e_1	e_2	e_3	e_4	e_5
h_1	1	1	0	1	0
h_2	0	1	0	1	0
h_3	1	1	0	0	0
h_4	0	0	1	0	0
h_5	0	0	1	0	1
h_6	1	1	0	1	0

3 A Soft Set Model on Equivalence Classes

There are many applications on information system including computation related to equivalence classes or attributes. The computation may be intersection, union of the sets of equivalence classes or dependency degree among attributes. It is inconvenient to execute these operations directly on the information system; therefore, based on the basic definition of soft set as in Definition 4, we construct a soft set model over equivalence classes to facilitate the computation related to equivalence classes and attributes. The soft set model is defined as follows.

Definition 5. *Given $S = (U, A, V, f)$ be an information system, let U/A denotes the set of all equivalence classes in the partitions U/a_i ($a_i \in A$ and $i = 1,2,\ldots,|A|$). Let $U' = U/A$ be the initial universe of objects, $E = U/A$ be the set of parameters, $P(U')$ denotes the power set of U', and define mapping $F: E \to P(U')$, we call the pair (F, E) a soft set model over equivalence classes.*

From this definition, for any equivalence class $e \in E$, $F(e) \subseteq U'$ is the set of equivalence classes which have some certain relations with *e*. By defining different mapping *F*, we can construct different soft sets to meet various requirements. Table 2 shows the tabular representation of the soft set over equivalence classes.

Table 2. Tabular representation of the soft set over equivalence classes

U'	e_1	e_2	e_i	…	e_m
x_1	$F(e_1)(x_1)$	$F(e_2)(x_1)$	$F(e_i)(x_1)$	…	$F(e_m)(x_1)$
x_2	$F(e_1)(x_2)$	$F(e_2)(x_2)$	$F(e_i)(x_2)$	…	$F(e_m)(x_2)$
x_i	$F(e_1)(x_i)$	$F(e_2)(x_i)$	$F(e_i)(x_i)$	…	$F(e_m)(x_i)$
…	…	…	…	…	…
x_m	$F(e_1)(x_m)$	$F(e_2)(x_m)$	$F(e_i)(x_m)$	…	$F(e_m)(x_m)$

Because $U'=E$, hence in Table 2 $e_i = x_i$ (i=1, …, m), where $m = |U/A| = |U'|$. $F(e_i)(x_i)$ equal 0 or 1.

Based on the proposed soft set model, in detail, we construct two soft sets to compute lower and upper approximation sets of an equivalence class or attribute with respect to other attributes in rough set.

We define two mappings F_1, F_2: $E \rightarrow P(U')$ as follows,

(1) F_1 : *Subsetof* (E)
(2) F_2 : *HasIntersectionWith* (E)

For any equivalence class $e \in E$, $F_1(e) \subseteq U'$ is the set of equivalence classes which are subsets of e, $F_2(e) \subseteq U'$ is the set of equivalence classes which have intersection with e. Having the two mappings, we can construct two soft sets (F_1, E) and (F_2, E). An illustrative example of the two soft sets is given in Example 2.

Example 2. We consider the information system as shown in Table 3, where $U = \{x_1 =1, x_2 =2, x_3 =3, x_4 =4, x_5 =5\}$, $A = \{a_1$ =Shape, a_2 =Color, a_3 =Area$\}$.

Table 3. An information system of objects' appearance

U	Shape	Color	Area
1	Circle	Red	Big
2	Circle	Red	Small
3	Triangle	Blue	Small
4	Triangle	Green	Small
5	Circle	Blue	Small

From Table 3, there are three partitions of U induced by indiscernibility relation on each attribute, i.e.

$$U/a_1 = \{\{1,2,5\},\{3,4\}\},\ U/a_2 = \{\{1,2\}, \{3,5\},\{4\}\},\ U/a_3 = \{\{1\},\{2,3,4,5\}\}.$$

We firstly construct the soft set (F_1, E). We have

$$U/A = \{\{1,2,5\},\{3,4\}, \{1,2\}, \{3,5\},\{4\}, \{1\},\{2,3,4,5\}\}.$$

Consequently,
$U' = E = U/A$, and then we can obtain

$$F_1(\{1,2,5\}) = Subsetof(\{1,2,5\}) =\{\{1,2,5\},\{1,2\}, \{1\}\};$$
$$F_1(\{3,4\}) = Subsetof(\{3,4\}) =\{\{3,4\},\{4\}\};$$
$$F_1(\{1,2\}) = Subsetof(\{1,2\}) =\{\{1,2\},\{1\}\};$$
$$F_1(\{3,5\}) = Subsetof(\{3,5\}) =\{\{3,5\}\};$$
$$F_1(\{4\}) = Subsetof(\{4\}) =\{\{4\}\};$$
$$F_1(\{1\}) = Subsetof(\{1\}) =\{\{1\}\};$$
$$F_1(\{2,3,4,5\}) = Subsetof(\{2,3,4,5\}) =\{\{3,4\},\{3,5\},\{4\},\{2,3,4,5\}\}.$$

The tabular representation of the soft set (F_1, E) is showed in Table 4.

Table 4. The tabular representation of the soft set (F_1, E)

U'	e_1 {1,2,5}	e_2 {3,4}	e_3 {1,2}	e_4 {3,5}	e_5 {4}	e_6 {1}	e_7{2,3,4,5}
x_1 {1,2,5}	1	0	0	0	0	0	0
x_2 {3,4}	0	1	0	0	0	0	1
x_3 {1,2}	1	0	1	0	0	0	0
x_4 {3,5}	0	0	0	1	0	0	1
x_5 {4}	0	1	0	0	1	0	1
x_6 {1}	1	0	1	0	0	1	0
x_7{2,3,4,5}	0	0	0	0	0	0	1

Similarly, we can construct the soft set (F_2, E). We have

U/A = {{1,2,5},{3,4}, {1,2}, {3,5},{4}, {1},{2,3,4,5}}.
$U' = E = U/A$, and then we can obtain
F_2({1,2,5})=*HasIntersectionWith* ({1,2,5}) ={{1,2,5},{1,2}, {3,5},{1},{2,3,4,5}};
F_2 ({3,4}) = *HasIntersectionWith* ({3,4}) ={{3,4},{3,5},{4},{2,3,4,5}};
F_2 ({1,2}) = *HasIntersectionWith* ({1,2}) ={{1,2,5},{1,2},{1},{2,3,4,5}};
F_2 ({3,5}) = *HasIntersectionWith* ({3,5}) ={{1,2,5},{3,4},{3,5},{2,3,4,5}};
F_2 ({4}) = *HasIntersectionWith* ({4}) ={{3,4},{4},{2,3,4,5}};
F_2 ({1}) = *HasIntersectionWith* ({1}) ={{1,2,5},{1,2},{1}};
F_2 ({2,3,4,5}) = *Subsetof*({2,3,4,5}) ={{1,2,5}, {3,4},{1,2},{3,5},{4},{2,3,4,5}}.

The tabular representation of the soft set (F_2, E) is showed in Table 5.

Table 5. The tabular representation of the soft set (F_2, E)

U'	e_1 {1,2,5}	e_2 {3,4}	e_3 {1,2}	e_4 {3,5}	e_5 {4}	e_6 {1}	e_7{2,3,4,5}
x_1 {1,2,5}	1	0	1	1	0	1	1
x_2 {3,4}	0	1	0	1	1	0	1
x_3 {1,2}	1	0	1	0	0	1	1
x_4 {3,5}	1	1	0	1	0	0	1
x_5 {4}	0	1	0	0	1	0	1
x_6 {1}	1	0	1	0	0	1	0
x_7{2,3,4,5}	1	1	1	1	1	0	1

Based on the soft sets (F_1, E) and (F_2, E), we build computational model for lower and upper approximation sets below.

The lower approximation of parameter e_j with respect to attribute a_i is defined as

$$\underline{a_i}(e_j)=\{x|x\in x_k \quad and F_1(e_j)(x_k)=1, k=D+1,...,D+|U/a_i|\} \tag{4}$$

where D refers to the total number of equivalence classes in the partitions U/a_l(l=1,…,i-1), namely $D=\sum_{l=1}^{i-1}|U/a_l|$, $x_k\in U'$ is one of the equivalence classes induced by U/a_i, $F_1(e_j)(x_k)\in\{0,1\}$ is the entry of the tabular representation of the soft set (F_1, E).

The cardinality of $\underline{a_i}(e_j)$ can be calculated as

$$\left|\underline{a_i}(e_j)\right| = \sum_{k=D+1}^{D+|U/a_i|} F_1(e_j)(x_k) \cdot |x_k| \tag{5}$$

The lower approximation of attribute a_j with respect to attribute a_i is defined as

$$\underline{a_i}(a_j) = \{x | x \in \underline{a_i}(e_k), k = D+1, ..., D+|U/a_j|\} \tag{6}$$

where $D = \sum_{l=1}^{j-1} |U/a_l|$.

The cardinality of $\underline{a_i}(a_j)$ can be calculated as

$$\left|\underline{a_i}(a_j)\right| = \sum_{k=D+1}^{D+|U/a_j|} \left|\underline{a_i}(e_k)\right| \tag{7}$$

Similarly, the upper approximation of parameter e_j with respect to attribute a_i is defined as

$$\overline{a_i}(e_j) = \{x | x \in x_k \quad and F_2(e_j)(x_k) = 1, k = D+1, ..., D+|U/a_i|\} \tag{8}$$

where D has the same meaning as in Eq. (4), namely $D = \sum_{l=1}^{i-1} |U/a_l|$, $F_2(e_j)(x_k) \in \{0,1\}$ is the entry of the tabular representation of the soft set (F_2, E).

The cardinality of $\overline{a_i}(e_j)$ can be calculated as

$$\left|\overline{a_i}(e_j)\right| = \sum_{k=D+1}^{D+|U/a_i|} F_2(e_j)(x_k) \cdot |x_k| \tag{9}$$

The upper approximation of attribute a_j with respect to attribute a_i is defined as

$$\overline{a_i}(a_j) = \{x | x \in \overline{a_i}(e_k), k = D+1, ..., D+|U/a_j|\} \tag{10}$$

where $D = \sum_{l=1}^{j-1} |U/a_l|$.

The cardinality of $\overline{a_i}(a_j)$ can be calculated as

$$\left|\overline{a_i}(a_j)\right| = \sum_{k=D+1}^{D+|U/a_j|} \left|\overline{a_i}(e_k)\right| \tag{11}$$

With the computational model, it is convenient to compute the lower and upper approximation of equivalence class or attribute with respect to other attributes. Let us reconsider Example 2. Suppose we are required to compute the cardinality of lower approximation of equivalence class e_1 = {1, 2, 5} with respect to attribute a_2, according to Eq. (5), we have

$$\left|\underline{a_2}(e_1)\right| = \sum_{k=3}^{5} F(e_1)(x_k)\cdot\left|x_k\right| = 1\times 2 + 0\times 2 + 0\times 1 = 2\,.$$

Furthermore, if we are required to compute the lower approximation of attribute a_1 with respect to attribute a_2, we have

$$\underline{a_2}(a_1) = \{x \,|\, x\in \underline{a_2}(e_1), \underline{a_2}(e_2)\} = \{1,2,4\}.$$

4 Application of the Proposed Model in Clustering Attribute Selection

Rough set theory based attribute selection clustering approaches for categorical data have attracted much attention in recent years. Mazlack et al. [7] proposed a technique using the average of the accuracy of approximation in the rough set theory called total roughness (TR). Parmar et al. [8] proposed a technique called Min-Min Roughness (MMR) for categorical data clustering. In selecting clustering attribute, the accuracy of approximation is measured using the well-known Marczeweski–Steinhaus metric applied to the lower and upper approximations of a subset of the universe in an information system [9]. Herawan et al. [10] proposed a new technique called maximum dependency attributes (MDA) for selecting clustering attribute, which is based on rough set theory by taking into account the dependency of attributes of the database. Compared to TR and MMR, MDA technique provides better performance.

In this section, first, we briefly introduce MDA technique, and then propose an algorithm based on the soft set constructed in Section 5, finally the comparison tests between soft set based algorithm and MDA technique are implemented on two UCI benchmark data sets.

4.1 MDA Technique

In information system $S = (U, A, V, f)$, given any attribute a_i, a_j, $k_{a_j}(a_i)$ refer to a_i depends on a_j in degree k, is obtained by Eq. (4) as follows

$$k_{a_j}(a_i) = \frac{\sum_{X\in U/a_i}\left|\underline{a_j}(X)\right|}{\left|U\right|} \tag{12}$$

Next, given m attributes, *Max-Dependency* (MD) of attribute $a_i (a_i \in A)$ is defined as

$$MD(a_i) = Max(k_{a_1}(a_i),\ldots,k_{a_j}(a_i),\ldots,k_{a_m}(a_i)) \tag{13}$$

where $a_i \neq a_j$, $1 \leq i, j \leq m$.

After obtaining the m values of $MD(a_i)$, $i = 1,2,...,m$. MDA technique selects the attribute with the maximum value of MD as clustering attribute, i.e.

$$MDA = Max(MD(a_1),...,MD(a_i),...,MD(a_m)) \tag{14}$$

4.2 An Algorithm Based on Soft Set (F1, E)

From Eq. (12), it can be seen that only lower approximation is required in MDA technique, and we can further simplify Eq. (12) as

$$k_{a_j}(a_i) = \frac{\left|\underline{a_j}(a_i)\right|}{|U|} \tag{15}$$

We propose the algorithm based on soft set (F_1, E) as follows,

Algorithm 1

1. *Compute the equivalence classes using the indiscernibility relation on each attribute.*
2. *Construct soft set (F_1, E).*
3. *Construct the tabular representation of the soft set (F_1, E).*
4. *Compute the cardinality of lower approximation of an attribute with respect to other attributes in terms of Eq. (7).*
5. *Compute the dependency of an attribute with respect to other attributes in terms of Eq. (15).*
6. *Select the attribute with the highest dependency as the clustering attribute.*

Let us reconsider the Example 2 following Algorithm 1. The first three steps has been shown in Example 2, we start with the fourth step, namely compute the cardinality of lower approximation in terms of Eq. (7). We can obtain

$$\left|\underline{a_2}(a_1)\right| = \left|\underline{a_2}(e_1)\right| + \left|\underline{a_2}(e_2)\right| = 2 + 1 = 3$$

$$\left|\underline{a_3}(a_1)\right| = \left|\underline{a_3}(e_1)\right| + \left|\underline{a_3}(e_2)\right| = 1 + 0 = 1$$

$$\left|\underline{a_1}(a_2)\right| = \left|\underline{a_1}(e_3)\right| + \left|\underline{a_1}(e_4)\right| + \left|\underline{a_1}(e_5)\right| = 0 + 0 + 0 = 0$$

$$\left|\underline{a_3}(a_2)\right| = \left|\underline{a_3}(e_3)\right| + \left|\underline{a_3}(e_4)\right| + \left|\underline{a_3}(e_5)\right| = 1 + 0 + 0 = 1$$

$$\left|\underline{a_1}(a_3)\right| = \left|\underline{a_1}(e_6)\right| + \left|\underline{a_1}(e_7)\right| = 0 + 2 = 2$$

$$\left|\underline{a_2}(a_3)\right| = \left|\underline{a_2}(e_6)\right| + \left|\underline{a_2}(e_7)\right| = 0 + 3 = 3$$

Next, we compute the dependency degree of an attribute with respect to other attributes in terms of Eq. (15). The results are summarized in Table 6.

Table 6. The degree of dependency of all attributes in Table 3

w.r.t	a_1	a_2	a_3
a_1	-	0.6	0.2
a_2	0	-	0.2
a_3	0.4	0.6	-

Taking a look at Table 6, attribute a_3 (Area) will be selected as clustering attribute.

5 Experimental Results

In order to test Algorithm 1 and compare with MDA technique in [10], we use two datasets Soybean and Zoo obtained from the benchmark UCI Machine Learning Repository [6]. The two methods are implemented in C++ language. They are sequentially executed on a PC with a processor Intel Core 2 Duo 2.0GHz. The main memory is 2 GB and the operating system is Widows XP Professional SP3.

5.1 Soybean Data Set

The Soybean data set contains 47 instances on diseases in soybeans. Each instance is described by 35 categorical attributes and can be classified as one of the four diseases namely, Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. The data set is comprised 17 objects for Phytophthora Rot disease and 10 objects for each of the remaining diseases. Fig.1 illustrates the executing time of selecting the clustering attribute.

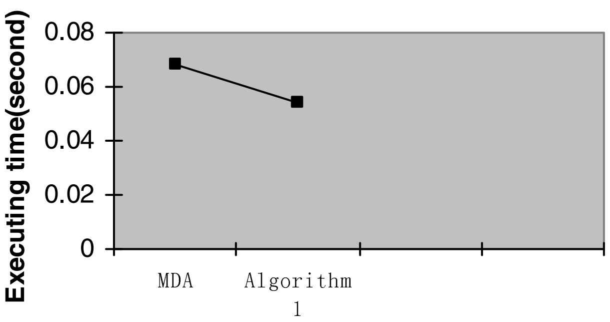


Fig. 1. The executing time of MDA and Algorithm 1on Soybean data set

5.2 Zoo Data Set

The Zoo data set contains 101 instances, where each instance represents information of an animal in terms of 16 categorical attributes. Each animal can be classified into seven classes namely, Mammal, Bird, Reptile, Fish, Amphibian, Insect, and Invertebrate. The data set is comprised 41 mammals, 20 birds, 5 reptiles, 13 fish, 4 amphibians, 8 insects and 10 invertebrates. Fig.2 illustrates the executing time of selecting the clustering attribute.

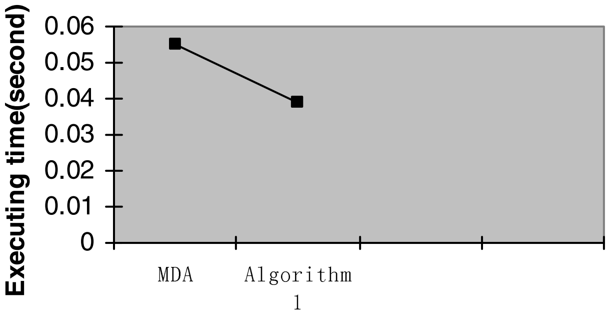


Fig. 2. The executing time of MDA and Algorithm 1on Zoo data set

From the above two experiments, it can be seen that Algorithm 1 improves the executing time of original MDA method.

6 Conclusions

In this paper, we present a soft set model on the set of equivalence classes in an information system. Based on the proposed model, in detail, we design two soft sets in order to obtain approximation sets of rough set. Furthermore, we make use of the proposed model to select clustering attribute for categorical data cluster and then a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Molodtsov, D.: Soft set theory_first results. Comput. Math. Appl. 37, 19–31 (1999)
2. Pawlak, Z.: Rough sets. International Journal Information Computer Science 11, 341–356 (1982)
3. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences: An International Journal 177(1), 3–27 (2007)
4. Feng, F., Li, C., Davvaz, B., Ali, M.I.: Soft sets combined with fuzzy sets and rough sets: a tentative approach. In: Soft Computing - A Fusion of Foundations, Methodologies and Applications, pp. 899–911. Springer, Heidelberg (2009)
5. Herawan, T., Deris, M.M.: A direct proof of every rough set is a soft set. In: Proceeding of the Third Asia International Conference on Modeling and Simulation, pp. 119–124 (2009)

6. UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRRepository.html
7. Mazlack, L.J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing clustering attributes. In: Proceedings of the ISCA 13th International Conference (CAINE 2000), pp. 1–6 (2000)
8. Parmar, D., Wu, T., Blackhurst, J.: MMR: an algorithm for clustering categorical data using rough set theory. Data and Knowledge Engineering 63, 879–893 (2007)
9. Yao, Y.Y.: Information granulation and rough set approximation. International Journal of Intelligent Systems 16(1), 87–104 (2001)
10. Herawan, T., Deris, M.M., Abawajy, J.H.: A rough set approach for selecting clustering attribute. Knowledge-Based Systems 23, 220–231 (2010)

Alternative Model for Extracting Multidimensional Data Based-On Comparative Dimension Reduction

Rahmat Widia Sembiring[1], Jasni Mohamad Zain[1], and Abdullah Embong[2]

[1] Faculty of Computer System and Software Engineering,
Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300, Kuantan, Pahang Darul Makmur, Malaysia
rahmatws@yahoo.com, jasni@ump.edu.my
[2] School of Computer Science, Universiti Sains Malaysia
11800 Minden, Pulau Pinang, Malaysia
ae@cs.usm.my

Abstract. In line with the technological developments, the current data tends to be multidimensional and high dimensional, which is more complex than conventional data and need dimension reduction. Dimension reduction is important in cluster analysis and creates a new representation for the data that is smaller in volume and has the same analytical results as the original representation. To obtain an efficient processing time while clustering and mitigate curse of dimensionality, a clustering process needs data reduction. This paper proposes an alternative model for extracting multidimensional data clustering based on comparative dimension reduction. We implemented five dimension reduction techniques such as ISOMAP (Isometric Feature Mapping), KernelPCA, LLE (Local Linear Embedded), Maximum Variance Unfolded (MVU), and Principal Component Analysis (PCA). The results show that dimension reductions significantly shorten processing time and increased performance of cluster. DBSCAN within Kernel PCA and Super Vector within Kernel PCA have highest cluster performance compared with cluster without dimension reduction.

Keywords: curse of dimensionality, dimension reduction, ISOMAP, KernelPCA, LLE, MVU, PCA, DBSCAN.

1 Introduction

In line with the technological developments, the current data tends to be multidimensional and high dimension, which is complex than conventional data. Many clustering algorithms have been proposed, but for multidimensional data and high dimensional data, conventional algorithms often produce clusters that are less meaningful. Furthermore, the use of multidimensional data will result in more noise, complex data, and the possibility of unconnected data entities. This problem can be solved by using clustering algorithm. Several clustering algorithms grouped into cell-based clustering, density based clustering, and clustering oriented. To obtain an efficient processing time to mitigate a curse of dimensionality while clustering, a clustering process needs data reduction.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 28–42, 2011.

Data reduction techniques create a new representation for the data that is smaller in volume and has the same analytical results as the original representation. There are various strategies for data reduction: aggregation, dimension reduction, data compression, discretization, and concept hierarchy [1]. Dimension reduction is a technique that is widely used for various applications to solve curse dimensionality.

Dimension reduction is important in cluster analysis, which not only makes the high dimensional data addressable and reduces the computational cost, but also can provide users with a clearer picture and visual examination of the data of interest [2]. Many emerging dimension reduction techniques proposed, such as Local Dimensionality Reduction (LDR). LDR tries to find local correlations in the data, and performs dimensionality reduction on the locally correlated clusters of data individually [3], where dimension reduction as a dynamic process adaptively adjusted and integrated with the clustering process [4].

Sufficient Dimensionality Reduction (SDR) is an iterative algorithm [5], which converges to a local minimum of $p^* = \arg \underset{\tilde{p} \in P\theta}{min} D_{KL}[p|\tilde{p}]$ and hence solves the Max-Min problem as well. A number of optimizations can solve this minimization problem, and reduction algorithm based on Bayesian inductive cognitive model used to decide which dimensions are advantageous [6]. Developing an effective and efficient clustering method to process multidimensional and high dimensional dataset is a challenging problem.

The main contribution of this paper is the development of an alternative model to extract data based on density connection and comparative dimension reduction technique. Results of extracting data implemented in DBSCAN cluster, and compare with other clustering method, such as Kernel K-Mean, Super Vector and Random Cluster. This paper is organized into a few sections. Section 2 will present the related work. Section 3 explains the materials and method. Section 4 elucidates the results followed by discussion in Section 5. Section 6 deals with the concluding remarks.

2 Related Work

Functions of data mining are association, correlation, prediction, clustering, classification, analysis, trends, outliers and deviation analysis, and similarity and dissimilarity analysis. Clustering technique is applied when there is no class to predict but rather when the instances divide into natural groups [7, 8]. Clustering for multidimensional data has many challenges. These are noise, complexity of data, data redundancy, and curse of dimensionality. To mitigate these problems dimension reduction needed. In statistics, dimension reduction is the process of reducing the number of random variables. The process classified into feature selection and feature extraction [9], and the taxonomy of dimension reduction problems [10] shown in Fig.1. Dimension reduction is the ability to identify a small number of important inputs (for predicting the target) from a much larger number of available inputs, and is effective in cases when there are more inputs than cases or observations.

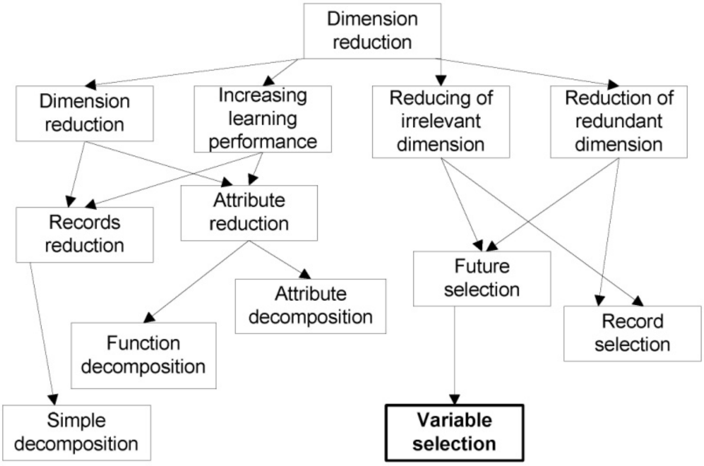


Fig. 1. Taxonomy of dimension reduction problem

Dimensionality reduction techniques have been a successful avenue for automatically extracting the latent concepts by removing the noise and reducing the complexity in processing the high dimensional data [11]. Maaten *et.al* proposed taxonomy dimension reduction technique as shown at Fig. 2, and found traditional dimensionality technique applied PCA and factor analysis, but this technique is unable to handle nonlinear data [12].

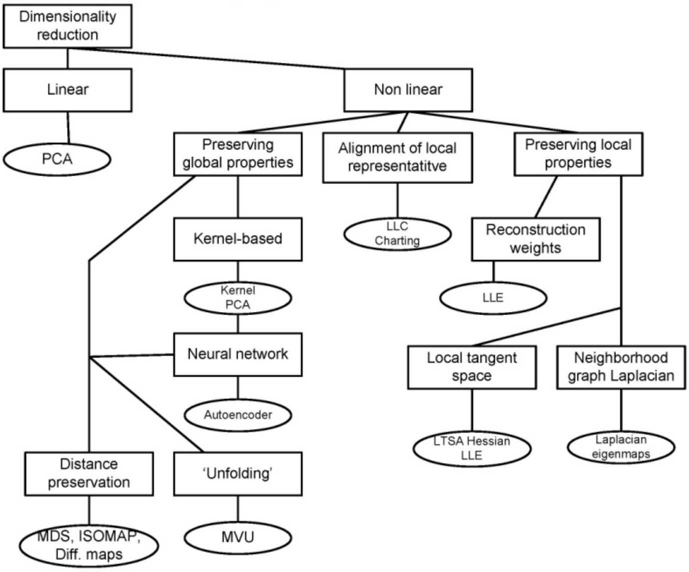


Fig. 2. Taxonomy of dimension reduction technique

The goals of dimension reduction methods are to reduce the number of predictor components and to help ensure that these components are independent. The method designed to provide a framework for interpretability of the results, and to find a mapping F that maps the input data from the space $\mathfrak{R}^d$ to lower dimension feature space $\mathfrak{R}^d$ denotes as $F(x)\colon \mathfrak{R}^d \rightarrow \mathfrak{R}^{\mathrm{d}'}$ [13, 14]. Dimension reduction techniques, such as principal component analysis (PCA) and partial least squares (PLS) can used to reduce the dimension of the microarray data before certain classifier is used [15].

We compared five dimension reduction techniques and embedded in 4 cluster techniques, these dimension reduction are:

A. ISOMAP

ISOMAP (Isometric Feature Mapping) is one of several widely used low-dimensional embedding methods, where geodesic distances on a weighted graph incorporated with the classical scaling. This approach combines the major algorithmic features of PCA and MDS [16, 17] computational efficiency, global optimality, and asymptotic convergence guarantees with the flexibility to learn a broad class of nonlinear manifolds. ISOMAP used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. ISOMAP is highly efficient and generally applicable to a broad range of data sources and dimensionalities [18]. ISOMAP Algorithm [16] provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbours on the manifold, such as the following phase:

a. Construct neighbourhood graph
   Define the graph G over all data points by connecting points j *and I* if as measured by $d_x\,(i, j)$, they are closer than e (*e-Isomap*), or if i is one of the K nearest neighbours of j (*K-Isomap*). Set edge lengths equal to $d_x(i, j)$. Determines which points are neighbours on the manifold M, based on the distances $dX\ (i, j)$ between pairs of points i, j in the input space X. These neighbourhood relations are represented as a weighted graph G over the data points, with edges of weight $dX(i, j)$ between neighbouring points.
b. Compute shortest paths
   Initialize $d_G(i, j) = d_x(i, j)$ if i, j are linked by an edge; $d_G(i, j) = \infty$ otherwise. Then for each value of $k = 1, 2, \ldots, N$ in turn, replace all entries $d_G(i, j)$ by $\min\{d_G(i, j), d_G(i,k) + d_G(k, j)\}$. The matrix of final values $D_G = \{d_G(i, j)\}$ will contain the shortest path distances between all pairs of points in G. ISOMAP estimates the geodesic distances $dM(i, j)$ between all pairs of points on the manifold M by computing their shortest path distances $dG(i, j)$ in the graph G.
c. Construct d-dimensional embedding
   Let λp be the *p-th* eigenvalue (in decreasing order) of the matrix $t(D_G)$, and $v'p$ be the *i-th* component of the *p-th* eigenvector. Then set the *p-th* component of the d-dimensional coordinate vector $\boldsymbol{y}_i$ equal to $\sqrt{}\lambda_p v'_p$. Final step applies classical MDS to the matrix of graph distances $DG\ 5\ \{dG(i, j)\}$, constructing an embedding of the data in a d-dimensional Euclidean space Y that best preserves the manifold's estimated intrinsic geometry. The coordinate vectors $\boldsymbol{y}_i$ for points in Y are chosen to minimize the cost function $E = \left\|\tau(D_{G)} - \ \tau(D_Y)\ \right\|L^2$, where D_Y denotes the matrix of Euclidean distance $\{d_Y(i,j) = \ \|y_i - \ y_j\,\|$ and $\|A\,\|L^2$ the matrix L^2

matrix norm $\sqrt{\sum_{i,j} A^2{}_{i,j}}$. The operator converts distances to inner products, which uniquely characterize the geometry of the data in a form that supports efficient optimization.

B. Kernel PCA

Kernel PCA is an extension of PCA [19], where PCA as a basis transformation to diagonalize an estimate of the covariance matrix of the data x_k, $k = 1,....,\ell$ $x_k \epsilon R^N$, $\sum_{k=1}^{l} x_k = 0$, defined as $C = \frac{1}{\ell}\sum_{j=1}^{y} x_j x_j^{\mathrm{T}}$. The Kernel PCA algorithm proceeds as follows:

a. Set a kernel mapping $k(x_m, x_n)$.
b. Count $\boldsymbol{K}$ based on $\{x_n, (n = 1, \ldots, N\ )\}$.
c. Find eigenvalue of $\boldsymbol{K}$ to get λ_i and a_i
d. For each given data point $\boldsymbol{X}$, find principal components in the feature space: $(f(x\ ) \,.\, \phi_i\ ) = \sum_{n=1}^{N} a_k^{(i)} k(x, x_n)$

In this paper, Gaussian kernel applied $k(x,y) = exp\left\{ -\frac{\left(-\|x-y\|^2\right)}{2\sigma^2} \right\}$

C. LLE

The LLE (Local Linear Embedded) algorithm based on simple geometric intuitions, where suppose the data consist of N real valued vectors $\vec{x}$ each of dimensionality, sampled from some smooth underlying manifold, the algorithm proposed [20]:

a. Compute the neighbours of each data point, $\overrightarrow{x_\imath}$
b. Compute the weight W_{ij} that best reconstruct each data point $\overrightarrow{X_\imath}$ from its neighbours, minimizing the cost in $\varepsilon(W) = \sum_i \left|\overrightarrow{x_\imath} - \sum_j W_{ij}\vec{X}_j\right|^2$ by constrained linear fits.
c. Compute the vectors $\vec{Y}_i$ best reconstructed by the weight W_{ij}, minimizing the quadratic form in $\Phi(\mathrm{Y}) = \sum_{\mathrm{i}} \left|\overrightarrow{\mathrm{Y}_\imath} - \sum_{\mathrm{i}} \mathrm{W}_{\mathrm{ij}}\vec{\mathrm{Y}}_{\mathrm{j}}\ \right|^2$

D. MVU

Maximum Variance Unfolded (MVU) is algorithms for nonlinear dimensionality reduction [21] map high dimensional inputs $\{\vec{x}_i\}_1^n = 1$ to low dimensional outputs $\{\vec{y}_i\}_1^n = 1$, where $\vec{x}_i \epsilon \Re^r, \vec{y}_i \epsilon \Re^r and\ r \ll d$. The reduced dimensionality r chosen to be as small as possible, yet sufficiently large to guarantee that the outputs $\vec{y}_i \epsilon \Re^r$ provide a faithful representation of the input s$\vec{x}_i \epsilon \Re^r$.

E. PCA

Principal Component Analysis (PCA) is a dimension reduction technique that uses variance as a measure of interestingness and finds orthogonal vectors (principal components) in the feature space that accounts for the most variance in the data [22]. Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis, first introduced by Pearson, and developed independently by Hotelling [23].

The advantages of PCA are identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. It is a powerful tool for analysing data by finding these patterns in the data. Then compress them by dimensions reduction without much loss of information [24]. Algorithm PCA [25] shown as follows:

a. Recover basis:
   Calculate $XX^T = \sum_{i=1}^{t} x_i x_i\ ^T$ and let U = eigenvectors of XX^T corresponding to the top d eigenvalues.
b. Encode training data:
   $Y = U^T X$ where Y is a $d\ x\ t$ matrix of encodings of the original data.
c. Reconstruct training data:
   $\hat{X} = UY = UU^T X$
d. Encode test example:
   $y = U^T > X$ where y is a d-dimensional encoding of x.
e. Reconstruct test example:
   $\hat{x} = U_y = UU^T x$

3 Material and Method

This study is designed to find the most efficient dimension reduction technique. In order to achieve this objective, we propose a model for efficiency of the cluster performed by first reducing the dimensions of datasets. There are five dimension reduction techniques tested in the proposed model, namely ISOMAP, KernelPCA, LLE, MVU, and PCA.

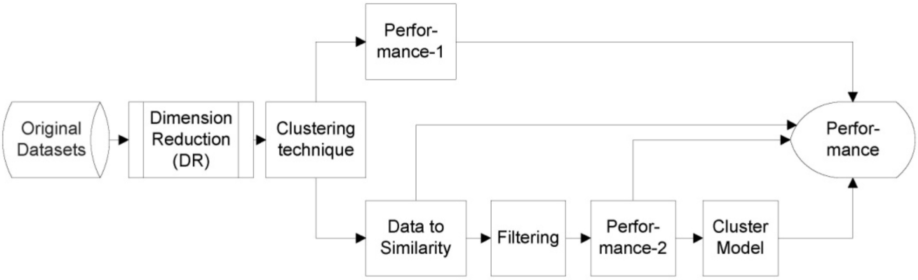


Fig. 3. Proposed model compared based on dimension reduction and DBSCAN clustering

Dimensions reduction result is processed into DBSCAN cluster technique. DBSCAN needs *ε (eps)* and the minimum number of points required to form a cluster (*minPts*) including mixed euclidean distance as distance measure. For the result of DBSCAN clustering using functional data to similarity, it calculates a similarity measure from the given data (attribute based), and another output of DBSCAN that is measured is *performance-1,* this simply provides the number of clusters as a value.

Result of *data to similarity* takes an exampleSet as input for filter examples and returns a new exampleSet including only the examples that fulfil a condition. By specifying an implementation of a condition, and a parameter string, arbitrary filters can be applied and directly derive a *performance-2* as measure from a specific data or statistics value, then process expectation maximum cluster with parameter *k*=2, *max runs*=5, *max optimization step*=100, *quality*=1.0E-10 and *install distribution=k-means* run.

4 Result

Testing of model performance was conducted on four datasets model; e-coli, iris, new machine cpu and thyroid. Dimension reduction used Isomap, Kernel PCA, LLE and MVU. Cluster technique used DBSCAN, Kernel K-Mean, Super Vector and Random Cluster. By using RapidMiner, we conducted the testing process without dimension reduction and clustering, and then compared with the results of clustering process using dimension reduction. Result of e-coli datasets process for processing time shown in Table 1a, and Table 1b for the performance of the cluster.

Table 1a. Processing time for e-coli datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	13	17	18	18
with Kernel PCA	14	24	15	14
with LLE	13	23	19	17
with MVU	13	18	15	15
with PCA	12	17	15	14
without dimension reduction	14	23	16	14

Table 1b. Performance of cluster for e-coli datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	99,4%	98,7%	99,4%	97,0%
with Kernel PCA	99,4%	98,7%	99,4%	97,1%
with LLE	99,4%	98,8%	99,4%	97,0%
with MVU	99,4%	98,7%	99,4%	97,0%
with PCA	99,4%	98,7%	99,4%	97,1%
without dimension reduction	99,4%	99,1%	99,4%	97,2%

Clustering process to iris datasets shown in Table 2a, for processing time and in Table 2b for the performance of the cluster.

Table 2a. Processing time for iris datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	11	6	6	6
with Kernel PCA	12	4	3	3
with LLE	11	10	7	7
with MVU	11	8	6	6
with PCA	10	5	4	4
without dimension reduction	11	8	7	7

Table 2b. Performance of cluster for iris datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	97,9%	93,5%	97,8%	91,2%
with Kernel PCA	98,7%	98,0%	98,7%	91,2%
with LLE	97,9%	95,6%	97,9%	91,2%
with MVU	97,9%	95,5%	97,8%	91,2%
with PCA	97,0%	98,0%	96,9%	93,9%
without dimension reduction	97,0%	98,0%	96,7%	93,9%

Machine cpu datasets consisting of 7 attributes and 209 samples clustered using the same method, and obtained the results shown in Table 3a, for processing time and Table 3b as a result of performance of the cluster.

Table 3a. Performance of cluster for machine cpu datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	11	3	4	5
with Kernel PCA	10	6	4	5
with LLE	8	4	5	5
with MVU	12	4	3	2
with PCA	11	7	9	7
without dimension reduction	13	15	22	19

Table 3b. Performance of cluster for machine cpu datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	98,6%	94,3%	33,3%	88,9%
with Kernel PCA	99,1%	66,7%	99,0%	95,4%
with LLE	97,2%	93,1%	97,2%	95,4%
with MVU	98,7%	99,4%	98,6%	88,9%
with PCA	40,0%	98,2%	0%	95,4%
without dimension reduction	99,5%	98,2%	99,5%	95,4%

Clustered result of new thyroid datasets shown in Table 4a for the processing time, and Table 4b for the performance of the cluster.

Table 4a. Performance of cluster for new thyroid datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	13	11	8	7
with Kernel PCA	17	7	9	9
with LLE	20	13	11	11
with MVU	17	13	12	9
with PCA	14	8	11	7
without dimension reduction	13	7	13	8

Table 4b. Performance of cluster for new thyroid datasets

Dimension reduction	Cluster Method			
	DBSCAN	Kernel K-Mean	Super Vector	Random Cluster
with ISOMAP	99,5%	96,9%	0%	95,5%
with Kernel PCA	99,1%	96,7%	99,1%	95,5%
with LLE	99,1%	98,9%	99,1%	95,5%
with MVU	98,7%	96,9%	0%	95,5%
with PCA	98,7%	96,7%	0%	95,5%
without dimension reduction	99,5%	98,3%	0%	95,6%

By implementing four different reduction techniques ISOMAP, KernelPCA, LLE, MVU, and PCA, and continuously applying the cluster method based on cluster density, We obtained results for the datasets of E.coli datasets. Some of the result We present at Fig. 4a-c. Fig. 4a is the result of the cluster with DBSCAN method that does not use a dimension reduction. Fig. 4b is the result of DBSCAN cluster method as well but first using dimension reduction. While Fig. 4c is the result of the cluster by using Super Vector and also use the dimension reduction.

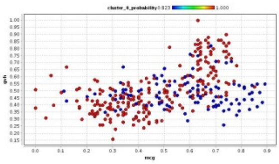


Fig. 4a. E-coli datasets based on DBSCAN without dimension reduction

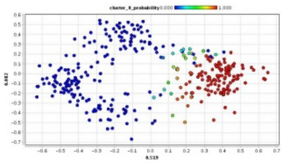


Fig. 4b. E-coli datasets based on DBSCAN and ISOMAP

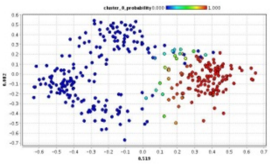


Fig. 4c. E-coli datasets based on Supervector and ISOMAP

For iris datasets consist of 4 attributes and 150 sample data, we implemented four different reduction techniques ISOMAP, KernelPCA, LLE, MVU, and PCA. We compared cluster result between process without dimension reduction and within dimension reduction. Some of the result present at Fig 5a-c. Fig. 5a is cluster result based on DBSCAN without dimension reduction. Fig. 5b is cluster result use DBSCAN within Kernel PCA as dimension reduction. This result similarly with Fig. 5c, cluster based on Random Cluster and Kernel PCA. Clustering process with dimension reduction create clearly different cluster (Fig. 5b and Fig. 5c).

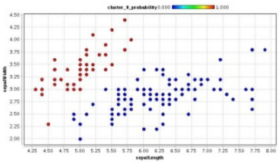


Fig. 5a. Iris datasets based on DBSCAN without dimension reduction

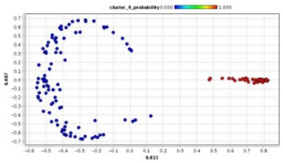


Fig. 5b. Iris datasets based on DBSCAN and Kernel PCA

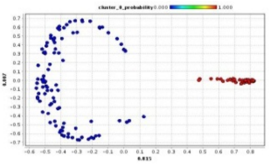


Fig. 5c. Iris datasets based on Random Cluster and Kernel PCA

The third was dataset tested is machine cpu. Some of the result we present at Fig 6a-c. In Fig. 6a shown cluster result based on DBSCAN without dimension reduction. Fig 6b. and Fig. 6c. was cluster result based on DBSCAN and Kernel K-Mean within using dimension reduction.

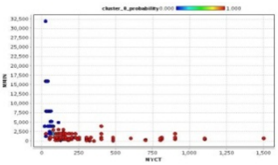


Fig. 6a. Machine cpu datasets based on DBSCAN without dimension reduction

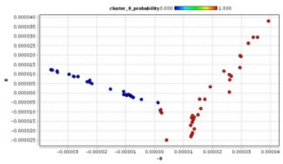


Fig. 6b. Machine cpu datasets based on DBSCAN and MVU

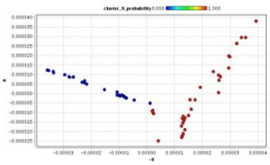


Fig. 6c. Machine cpu datasets based on Kernel K-Mean and MVU

Using same dimension reduction techniques, we clustered new thyroid datasets. We obtained results of DBSCAN without dimension reduction in Fig. 7a. While DBSCAN with dimension reduction using LLE has result in Fig. 7b. Cluster based Super Vector using LLE shown in Fig. 7c, we can see clustering process with dimension reduction create clearly different cluster (Fig. 7b. and Fig 7c.).

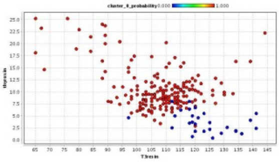


Fig. 7a. Machine cpu datasets based on DBSCAN without dimension reduction

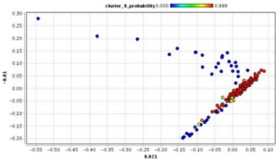


Fig. 7b. Machine cpu datasets based on DBSCAN and LLE

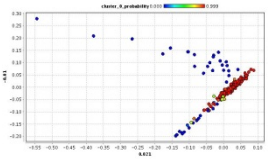


Fig. 7c. Machine cpu datasets based on Super Vector and LLE

Each cluster process, especially ahead of determined value of □=*1*, and the value *MinPts*=*5*, while the number of clusters (*k*=*2*) that will be produced was also determined before.

5 Discussion

Dimension reduction before clustering process is to obtain efficient processing time and increase accuracy of cluster performance. Based on results in previous section, dimension reduction can shorten processing time. Fig. 8a shows DBSCAN with PCA has lowest processing time. For iris datasets, we also found dimension reduction could shorten processing time. In Fig. 8b. Super Vector and Random Cluster within Kernel PCA has lowest processing time.

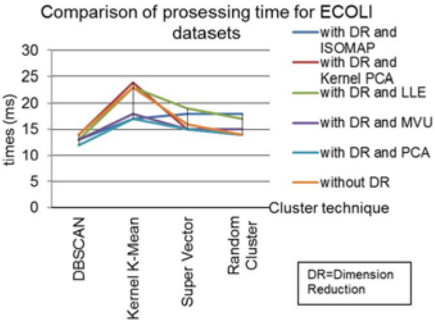


Fig. 8a. Performance of processing time for e-coli datasets using different dimension reduction technique and cluster technique

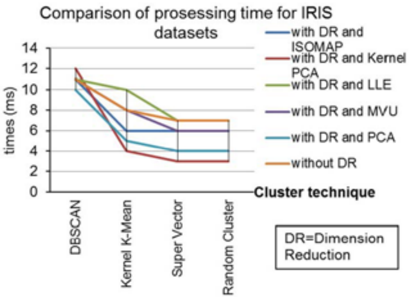


Fig. 8b. Performance of processing time for iris datasets using different dimension reduction technique and cluster technique

For machine cpu datasets, we found dimension reduction for Super Vector and Random Cluster within Kernel ISOMAP has lowest processing time (Fig. 8c). For new thyroid datasets, we found dimension reduction for Kernel K-Mean within Kernel PCA and Random Cluster within Kernel ISOMAP has lowest processing time (Fig. 8d).

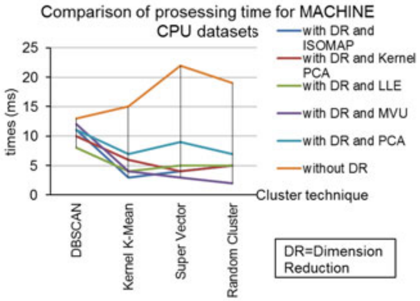


Fig. 8c. Performance of processing time for machine cpu dataset using different dimension reduction technique and cluster technique

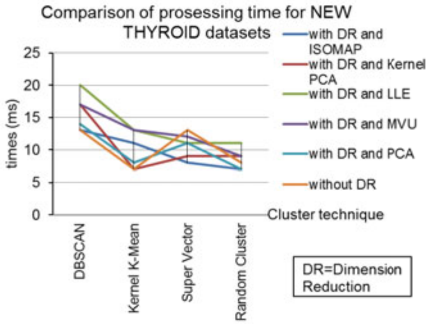


Fig. 8d. Performance of processing time for new thyroid datasets using different dimension reduction technique and cluster technique

Another evaluation for model implementation is comparison of cluster performance. In general dimension reduction increased cluster performance. For ecoli datasets we found Super Vector ISOMAP has highest cluster performance (Fig. 9a.). For iris dataset we found DBSCAN within Kernel PCA and Super Vector within Kernel PCA have highest cluster performance compared with cluster without dimension reduction (Fig. 9b.).

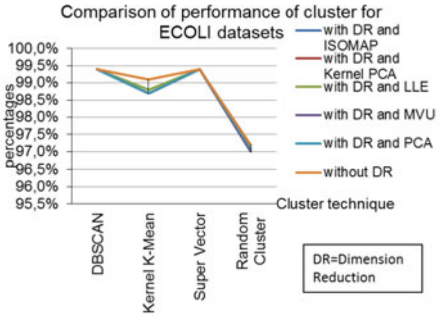


Fig. 9a. Performance of cluster for e-coli datasets using different dimension reduction technique

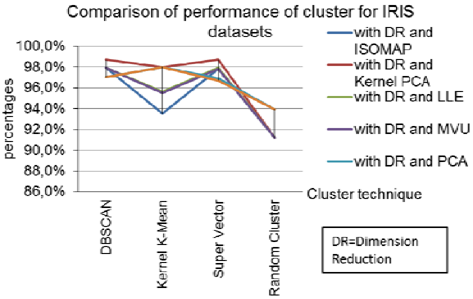


Fig. 9b. Performance of cluster for iris datasets using different dimension reduction technique

For machine cpu dataset in general cluster process without dimension reduction have highest cluster performance. Datasets, only Kernel K-Mean within PCA has cluster performance equal to cluster without dimension reduction (Fig. 9c.). For new thyroid dataset, we found Kernel K-Mean within LLE and Super Vector within LLE has highest cluster performance (Fig. 9d.).

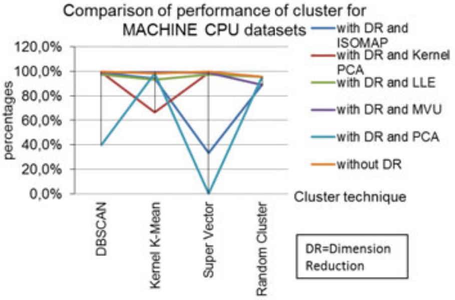


Fig. 9c. Performance of cluster for machine cpu datasets using different dimension reduction technique

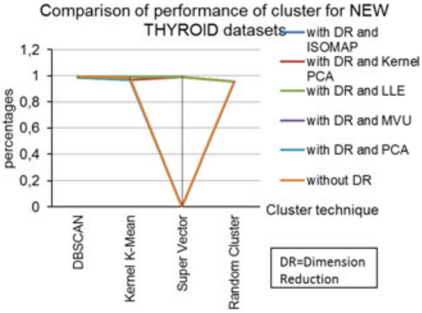


Fig. 9d. Performance of processing time for new thyroid datasets using different dimension reduction technique and cluster technique

6 Conclusion

The discussion above has shown that applying a dimension reduction technique will shorten the processing time.

Dimension reduction before clustering process is to obtain efficient processing time and increase accuracy of cluster performance. DBSCAN with PCA has lowest processing time for e-coli datasets. Super Vector and Random Cluster within Kernel PCA has lowest processing time for iris datasets. For machine cpu datasets, we found dimension reduction for Super Vector and Random Cluster within Kernel ISOMAP has lowest processing time. For new thyroid datasets, we found dimension reduction for Kernel K-Mean within Kernel PCA and Random Cluster within Kernel ISOMAP has lowest processing time.

In general, dimension reduction shows an increased cluster performance. For e-coli datasets, we found Super Vector ISOMAP has highest cluster performance. For iris datasets, we found DBSCAN within Kernel PCA and Super Vector within Kernel PCA have highest cluster performance compared with cluster without dimension reduction. For machine cpu dataset, in general cluster process without dimension reduction have highest cluster performance. For new thyroid datasets, we found Kernel K-Mean within LLE and Super Vector within LLE show the highest cluster performance.

References

1. Maimon, O., Rokach, L.: Decomposition Methodology For Knowledge Discovery And Data Mining, pp. 253–255. World Scientific Publishing Co, Pte, Ltd., Danvers (2005)
2. Fodor, I.K.: A Survey of Dimension Reduction Techniques. LLNL Technical Report, UCRL-ID-148494, p.1–18 (2002)
3. Chakrabarti, K., Mehrotra, S.: Local Dimensionality Reduction: A New Approach To Indexing High Dimensional Space. In: Proceeding of the 26th VLDB Conference, Cairo, Egypt, pp. 89–100 (2000)
4. Ding, C., He, X., Zha, H., Simon, H.: Adaptive Dimension Reduction For Clustering High Dimensional Data, pp. 1–8. Lawrence Berkeley National Laboratory (2002)
5. Globerson, A., Tishby, N.: Sufficient Dimensionality Reduction. Journal of Machine Learning, 1307–1331 (2003)
6. Jin, L., Wan, W., Wu, Y., Cui, B., Yu, X., Wu, Y.: A Robust High-Dimensional Data Reduction Method. The International Journal of Virtual Reality 9(1), 55–60 (2010)
7. Sembiring, R.W., Zain, J.M., Embong, A.: Clustering High Dimensional Data Using Subspace And Projected Clustering Algorithm. International Journal of Computer Science & Information Technology (IJCSIT) 2(4), 162–170 (2010)
8. Sembiring, R.W., Zain, J.M.: Cluster Evaluation Of Density Based Subspace Clustering. Journal of Computing 2(11), 14–19 (2010)
9. Nisbet, R., Elder, J., Miner, G.: Statistical Analysis & Data Mining Application, pp. 111–269. Elsevier Inc., California (2009)
10. Maimon, O., Rokach, L.: Data Mining And Knowledge Discovery Handbook, pp. 94–97. Springer Science+Business Media Inc., Heidelberg (2005)
11. Kumar, C.A.: Analysis Of Unsupervised Dimensionality Reduction Technique. ComSIS 6(2), 218–227 (2009)
12. van der Maaten, L. J. P., Postma, E.O., van den Herik, H.J.: Dimensionality Reduction: A Comparative Review. Published online, pp. 1–22 (2008), `http://www.cs.unimaas.nl/l.vandermaaten/dr/dimensionreduction_draft.pdf`
13. Xu, R., Wunsch II, D.C.: Clustering, pp. 237–239. John Wiley & Sons, Inc., New Jersey (2009)
14. Larose, D.T.: Data Mining Methods And Models, pp. 1–15. John Wiley & Sons Inc., New Jersey (2006)
15. Wang, J.: Encyclopedia Of Data Warehousing And Data Mining, p. 812. Idea Group Reference, Hershey (2006)
16. Tenenbaum, J., De Silva, V., Langford, J.C.: A Global Geometric Framework For Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
17. Mukund, B.: The Isomap Algorithm and Topological Scaling. Science 295, 7 (2002)
18. `http://www.wikipedia.com`
19. Schölkopf, B., Smola, A., Muller, K.R.: Non Linear Kernel Principal Component Analysis. Vision And Learning, Neural Computation 10, 1299–1319 (1998)
20. Saul, L.K.: An Introduction To Locally Linear Embedding, AT&T Labs–Research pp. 1–13 (2000), `http://www.cs.nyu.edu/~roweis/lle/papers/lleintroa4.pdf`
21. Weinberger, K.Q., Saul, L.K.: An Introduction To Nonlinear Dimensionality Reduction By Maximum Variance Unfolding. In: AAAI 2006 Proceedings of The 21st National Conference On Artificial Intelligence, vol. 2, pp. 1683–1686 (2006)

22. Poncelet, P., Teisseire, M., Masseglia, F.: Data Mining Patterns: New Methods And Application, Information Science Reference, Hershey PA, pp. 120–121 (2008)
23. Jolliffe, I.T.: Principal Component Analysis, pp. 7–26. Springer, New York (2002)
24. Smith, L.I.: A Tutorial On Principal Component Analysis (2002), http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdfp.12-16
25. Ghodsi, A.: Dimensionality Reduction, A Short Tutorial, Technical Report 2006-14, Department of Statistics and Actuarial Science, University of Waterloo, pp. 5–6 (2006)

Knowledge Taxonomy for Developing Organizational Memory System (OMS) for Public Institutions of Higher Learning (IHL) in Malaysia

Suzana Basaruddin[1] and Haryani Haron[2]

[1] Universiti Selangor, Faculty of Information Technology Industry, Bestari Jaya Campus,
45600 Bestari Jaya, Selangor, Malaysia
suzana_b@unisel.edu.my
[2] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40000 Shah Alam, Selangor, Malaysia
haryani@tmsk.uitm.edu.my

Abstract. Knowledge in organization, supported by appropriate technology will produce organizational memories that provide necessary support for organization to learn and grow. As the scope of organizational memory in one organization is too broad, it is necessary to classify them into acceptable narrower view. Using organization memory computer base literature related, this study found some basic understanding of organization memory system, and produce knowledge taxonomy and metadata as a base to develop organizational memory system in IHL; Malaysia scenario. The taxonomy and metadata is developed base on pertinent literature of organizational memory system model, organizational memory sources types and knowledge management technology. Further study can be conducted to validate and verify the developed taxonomy.

Keywords: Knowledge, taxonomy, organizational memory, corporate memory, knowledge management.

1 Introduction

The usage of Information and Communication Technology (ICT) has become the norm in our daily activities and businesses. Organizations' processes including registration, applications submission and operations, marketing and customer services have been computerized. ICT enable data, information, knowledge, and wisdom of organization to be shared. Organizational Memory (OM) is being treated as human memory that consists of history, good and bad experiences of the organization. It is one of way to manage knowledge at organization level. Learning curve is shortened and organization becomes mature with appropriate OM. As organizations realize the importance of OM, there has been so many knowledge management system developed to support organizational memory. Knowledge applications blooms and people are having more access to knowledge but some knowledge repositories remains in their own group or team instead of being shared across the organization. The reality on computerized system is that they have not being shared across organization. Finally, when the knowledge cannot be used as references and lessons learned, organization failed

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 43–57, 2011.

to become learning organization; a reflection of failure in organizational memory concept.

A classification principle of Organization Memory System (OMS) will bring new concrete views related to roles of Knowledge Management System (KMS) in organization; in particular focus to the cross functional knowledge sharing. This study aims to develop taxonomy for OMS. The taxonomy serves as a benchmark on organization for producing framework of OMS in IHL organization in Malaysia. This study differ from other framework that integrating information system (IS) in their knowledge production environment.

This paper is organized as follows. First section is the Introduction (Section 1). Section 2 covers theoretical framework from literature related to organizational memory and taxonomy before presenting method for developing the taxonomy (Section 3). Next is developing OMS taxonomy (Section 4) and finally summarizing results and providing suggestions for future research (section 5).

2 Theoretical Framework

2.1 Knowledge in Organization

The last twenty years have seen a major shift in world wide access to codified knowledge [1]. Human kind able to capture most of knowledge into computer system that provides access to its authorized user whenever necessary. To be effective, organizations must exploit example of best practice, improve their efficiency and contribute to their overall organizational learning [2]. Knowledge resides in various places and format such as database, knowledge bases, filing cabinets and peoples' head and is distributed right across the enterprise [3]. [4] emphasized that organization which adopts image of memory as a guide to the way in which it operates needs to address its ability to learn in real time: for its members to be able to make sense of a changing environment and for their sense making to spread rapidly throughout the organization or at least to those part where it is needed. Knowledge generated by organization activities often stays within a laboratory or research team and rarely crosses disciplinary boundaries, in most setting it is resides in individual knowledge cluster; so it is the challenge of institutional leadership to motivate their staff to share in many ways [5]. Decisions and reason could not be concluded as a whole because some of the knowledge remains in its save repositories. OM plays a role to centralize repositories in organizations so that there is only one repository in one organization contain the most update and reliable knowledge cross organization.

For a corporate or organizational memory to live and grow with organization, members of organization must support the Knowledge Management (KM) processes that are knowledge creation, organize, refine and transfer [6].Team learning is the ability of each team member to envision an alignment, between individual ability and the team's vision to produce greater results than can otherwise be delivered [7]. In one organization, knowledge gained by individual and teams has to be shared. Internal knowledge exchange process is actually a learning process to the organization. KM should archive the dynamic management of process creating knowledge out of knowledge rather than static management of information on existing knowledge [8].

For an organization to become one learning entity, it has to overcome barriers of individual/team learning; able to arrive at common understanding of company purpose and known organization problems; and exhibit a certain level of error tolerance (i.e incorrect learning or learning from bad/critical experiences) [9]. Organizational memory must not include only characteristic of just certain individuals, organizational memory is independent from any members. Organizational memory is knowledge from the past where experience mapped on present activities, thus resulting in higher of organizational effectiveness [10]. Lack of memories due to staff replacement can cause "corporate amnesia" [11].

2.2 Organizational Memory System (OMS)

OM term has been introduced since early 80s. Since then the term has been used in various research as listed in Table 1.

Table 1. Evolution of OM terms [12] & [13]

Author/s	Year	Terms used
Duncan and Weiss	1979	Corporate, organizational, enterprise wide knowledge base
Hedberg	1981	OM
Pautzke	1989	Corporate, organizational, enterprise wide knowledge base
Walsh and Ungson	1991	OM
Pralahad and Hamel	1994	Corporate Knowledge or Corporate Genetic
Rao and Goldmann	1995	OM, Corporate Memory
Rose et al.	1998	Corporate Memory
Annie	1999	Corporate Memory
Dieng et al.	1999	Corporate Memory (CM) (noncomputational CM, document-based CM, knowledgebased CM, case-based CM and distributed CM)

Studies on OM has been intensified since 1980's. Researchers have used the term OM interchangeable with Corporate Memory (CM), Corporate Knowledge (CK) and Enterprise Knowledge Base (EKB). Previous studies focus on the concept of OM and

OMIS. Those researchers concentrate on the concept of OM systems for solving information system and knowledge management problems. From definitions provided by those pioneers in the field, it can be concluded that OM is enterprise wide knowledge management focusing on centralize, transparent and cross departmental access. Among all researchers, [14] has made an effort to categorized sources of OMS and proposed six types of OM depicts in table 2.

Table 2. Types of OM [14]

No	Type of OM	Creation of OM
1	Non computational OM	Paper based documents on knowledge that had never been elicited previously
2	Document based OM	All existing docs of firm can constitute the OM – not well indexed
3	Knowledge based OM	Knowledge engineering for building OM
4	Case based OM	Collection of pasts experiences (successes or failures) that can be represented explicitly
5	Construction of distributed OM	Supporting collaboration and knowledge sharing between several groups of people in organization
6	Combination of several techniques	Both informal knowledge (such as documents) and formal knowledge (such as knowledge explicitly). represented in a knowledge base)

The categorization of OMS source proposed by [14] is adopted in this study because it reflects the context of knowledge resources and information in IHL.

In developing the taxonomy for this research, a critical analysis related to OMS are conducted. A compilation of studies from [12], [13] and [15] on models and frameworks related to OMS are listed in Table 3.

Table 3. Studies of OMS

No	Year	Outcome of Research *(OMS name)*	Short Description of Outcome *(OMS components)*
1	1986	Transactive Memory System	i) Individual System ii) External Memory iii) Transactive memory
2	1989	Organizational Knowledge Model	Knowledge accessible and not accessible within organization projected through overlapped circle.
3	1991	Organizational IT Memory framework called Bins.	i) Individual culture ii) Culture iii) Transformation iv) Structure v) Ecology vi) External Environment

Table 3. (*continued*)

4	1992	OMIS Success Model	i) System Quality, ii) Information Quality, iii) Success Measure in Terms of Usage, iv) Individual Impact, v) Organizational Impact
5	1995	Five Metadata types *(TeamBox)*	i) Meta-data ii) Structured data iii) Semistructured Data iv) Unstructured Data v) Temporal Data
6	1995	Two layers OMIS framework -IT and Non IT- Improved previous study	i) Effectiveness functions (integration, adaptation, goal attainment, pattern maintenance) ii) Mnemonic functions (knowledge acquisition, retention, maintenance, search and retrieval)
7	1996	FAQ in tree structure Information System *(Answer Garden)*	Database with answers to FAQ supported by expert.
8	1996	Knowledge Construction in material practice *(SHAMAN)*	Foster sharing of knowledge and experience
9	1996	OM informal knowledge *(QuestMap)*	OM creation and usage must occur in work group activity. Conversations in meetings represented in a graphical map format. Consist of hypertext, groupware and rhetorical method.
10	1997	Three components of OMIS	i) Paper Documents, ii) Computer Documents, iii) Self Memory
11	1997	Managing discussions *(Virtual Participant System)*	Developed for computer supported collaborative learning environment.
12	1998	Closed user group *(Knowledge Sharing Environment)*	Agent filters information obtained from database according to the profile defined by the user.
13	1998	OMIS Success Model - Five success factor/block-Improved previous study	i) System Quality, ii) Information Quality, iii) Success Measure in Terms of Usage, iv) Individual Impact, v) Organizational Impact

Table 3. (*continued*)

14	1998	OM model	i) Capture ii) Store iii) Disseminate iv) Facilitate use
15	1998	Organizational memory model	i) People (role, culture, position, social network), ii) Text (table, document), iii) Multimedia (image, audio, graphic, video), iv) Model, v) Knowledge
16	1999	Organizational Memory framework -IT and Non-IT- Improved from previous study	i) Individual culture ii) Culture iii) Transformation iv) Structure v) Ecology vi) External Environment vii) Non-IT Record viii) Files Elements
17	2000	OM of cooperative work *(WSISCO, OMUSISCO and PRIME)* using OMIS model	Web based system to manage structured discussion on the agenda item to be included in a decision making meeting. Consist of hypertext, groupware, rhetorical method, retrieval information system and privacy.
18	2000	Internet Based Knowledge Management	i) Acquire ii) Organize iii) Distribute
19	2004	Corporate Knowledge Management based on 6 types of Corporate Memory	6 types of Corporate Memory (Non computational, document based, knowledge based, case based, construction of distributed and combination of several techniques) Individual memory, project memory and managerial memory
20	2005	IHL Framework of KM system based on KM characteristics	5 components of KM framework (Psychological, culture, process, functionality, architecture)
21	2005	OM Knowledge Sharing	Formal and informal knowledge forming knowledge repositories for OM

Table 3. (*continued*)

22	2006	IHL OMIS implementation using previous study*(MemorIS)*	Evaluation of OMIS implementation for academic management: i) Staff characteristic ii) Work culture iii) Meta-memory as repository
23	2009	IHL OMIS implementation using previous study *(MemorIS)*	3 databases (user, programs & checklist), 5 success factors (system quality, information quality, usage success measure, individual impact & organization impact), metadata curriculum

The compilation of studies related to OMS provides various views of OMS done by previous researchers from 1986 to 2009. The 23 years of studies shows that 61% of the researchers studied on components of OMS. The rest came out with identification of success factors, processes and database manipulation to represent their view of OMS. The analysis of their researches falls into 4 categories as describe in Table 4.

Table 4. Categorization of OMS previous researches

No	Category of previous OMS research	Research numbered
1	Components	1,2,3,5,6,9,10,15,16,17,19,20, 21 and 22
2	Success factor	4, 13, 23
3	Process	6, 8,14,18
4	Database manipulation	7,12, 23

Previous researchers categorized the outcome of their study to OMS components (14 researches), success factors (3 researches) and processes (4 researches). Researches numbered 7, 12 and 23 manipulated database for problem solving. Out of 23 previous studies, 7 of the researchers proposed framework or model that view knowledge at individual level (researches numbered 1, 3, 4, 10, 13, 16 and 19). Metadata is included as main components in OMS by researches numbered 5, 22 and 23. The OMS initiation is viewed on individual system as suggested by the 7 previous researches. This study agreed on the importance of metadata and therefore incorporates metadata as one of main components in the final outcome. Components category derivations are varies as listed in Table 5.

OMS derivation components decision in this study is based from analysis in Table 5. It is obvious that most researchers derived OMS components from combination of documents and memories. This common understanding among previous researchers found in this study proven that combination of documents and memories are the major components of knowledge in most organizations. The OMS for IHL incorporates OMS model from literature analysis discussed in Table 2 until Table 5. The implementation of KM; can be viewed from two main areas that are KM techniques and KM technologies [16] as listed in Table 6.

Table 5. Analysis of knowledge components derivation for OMS defined by previous researcher

No	Knowledge components derivation	Research numbered
1	Surrounding of work place	3, 16, 20 & 22
2	Categorization of data	5
3	Effectiveness and mnemonic functions	6
4	Combination of documents and memories	1, 2, 9, 10, 15, 17, 19, 21

Table 6. KM Techniques and KM Technologies [16]

KM techniques	KM Technologies
1. Mentorship programs, 2. After action reviews or project summaries, 3. Regular intra office or intra division meetings, 4. Storytelling, 5. Communities of practice and 6. Centers of excellence	1. knowledge storage tools, 2. search and retrieval tool, 3. collaboration tools and 4. communication tools under KM technology

[16] claimed that KM techniques are the most effective techniques at capturing tacit knowledge, whereas KM technology is best at capturing explicit knowledge. KM approaches may fail when stakeholders rely on inadequate technology [17]. Thus KM technology plays important roles in organization as enabler to complete KM processes in organization from knowledge creations until knowledge transfers so that knowledge is available and reliable for future use (organizational memory). KM technology also includes knowledge related contents [18]. Selecting and using appropriate technology will determine success of the knowledge management implementation [19]. This is the main reason why this study included KM technology into the OMS taxonomy. KM technologies in this study is derived from isolated researchers' observation, of technology available in IHL, and then mapped into OMS taxonomy level 5. OMS should be a center of knowledge in organization; where KM technology plays its role as a tool to share knowledge.

2.3 IHL Scenario

Developing OMS for public IHL has its own challenges. OMS facilitates learning for an organization specifically for public IHL in Malaysia. System flow relevant to daily activities and IT artifacts are the main focus in this study. OM creation and usage must not be considered as an isolated activity but as daily practices. People should be willing to contribute for OM retention when they get direct benefits from the system. For this reason, this research will use the Key Performance Indicators (KPI) report as a force for people participations. This decision is made because it will drive people to use the system and raise chances of success implementation. Humans have limited ability of memory and have limited capacities to perform in their job responsibility [15]. Processes are the main component in delivering organizational goals. Thus, any approach that is not associated with processes will tend to fail or to be perceived as failures. Technology cannot be considered alone, it is limited to supporting humans because of its variable accuracy levels when performing simple mundane human tasks [17]. KM design without inputs from all stakeholders is one of the factors why KM approaches fail [12]. KPI report is the organization process and as part of organization culture to support the organization process. Employee is willing to support this process since it will return benefits to them. After all KPI is the widely used tool for employee performance evaluations in Malaysia. Information systems research is important here because it specifically deals with how the artifacts of IT interact with individuals in producing organizational outcomes [20]. [21] in his research illustrates how public sector organizations can avoid the ‘‘great trap in knowledge management” by focusing on designing IT artifacts to make explicit the tacit knowledge from people, and not in the information contained in document repositories.

[22] notes that many universities in Malaysia were not optimizing knowledge (shared and reused to generate new knowledge). [23] came out with taxonomy of IHL k-portal. This is one initiative to identify initial knowledge sharing platform for organization to do retention of their corporate memory. Their study covers explicit information on selected IHL portal. [24] founded three dimensions of tacit knowledge among academician that are intellectual affirmation, self and social tacit knowledge, emerged when academician faced challenges. IHL's organization knowledge is scattered, unstructured and unorganized. Knowledge in organization is stored in various forms of sources such as in repositories, databases, data warehouse, organization documents and all digital communication medias such as email, video and audio. In academic sector, thesis, curriculum, subject registration are among many other academic artifact need to be managed wisely [15] While [22] suggested that researchers of Computer and Information Sciences work out the needs for which ICT system are adopted by the very remarkable user group. [5] proposed “Content must be described and accessed in standardized and interoperable ways”. Both the suggestions lead to developing taxonomy in a specific domain. Suitable taxonomies play an important role in research and management because the classification of objects helps researchers and practitioners understand and analyze complex domains [26]. Any organization that needs to make significant volumes of information available in an efficient and consistent way to its customers, partners or employees, need to understand the value of a serious approach to taxonomy management [27].

2.4 Taxonomy

Science of classification or taxonomy is derived from Greek words (tassein + nomos) and Carl Linnaeus (1707 – 1778) was the first to use the idea of taxonomy to classify the natural world [28] At its simplest, taxonomy is a rule-driven hierarchical organization of categories used for classification purposes with the appropriate subject headings and descriptors. [28] points out that the components of developing corporate taxonomies are best understood by reviewing both the research literature as well as industry efforts; namely, key standards, metadata, classifiers, expertise locators and taxonomy tools available for automating what could be a massive undertaking.

Discussion on taxonomies is therefore not complete without an understanding of some of the fundamental vocabularies of knowledge organization and how they are derived.

When dealing with bases of explicit knowledge stored in electronic format, any taxonomy utilized is tightly coupled with the body of metadata utilized to define, identify, point, describe and characterize the contents of the knowledge base [29]. [28] has come out with Functionalities of Taxonomy Builders and Classifiers (Table 7). It projected all the main components related to taxonomy.

Table 7. Functionalities of Taxonomy Builders and Classifiers [28]

No	Item	Method
1	Classification Methods	Rule-based, Training Sets, Statistical Clustering, Manual
2	Classification Technologies	Linguistic Analysis, Neural Network, Bayesian Analysis, Pattern Analysis/ Matching, K-Nearest Neighbours, Support Vector Machine, XML Technology, Others
3	Approaches to Taxonomy Building	Manual, Automatic, Hybrid
4	Visualization Tools Used	Tree/Node, Map, Star, Folder, None
5	Depth of The Taxonomy	> 3 levels
6	Taxonomy Maintenance	Add/Create, Modify/Rename, Delete, Reorganization/Re-categorization, View/Print
7	Import/Export Formats Support	Text file, XML format, RDBS, Excel file, Others
8	Document Formats Support	HTML, MS Office document, ASCII/text file, Adobe PDF, E-mail, Others
9	Personalization	Personalized View, Alerting/Subscribing
10	Product integration	Search Tools, Administration Tools, Portals, Legacy Applications (e.g. CRM)
11	Industry-specific Taxonomy	Access Points to the Information: Browse Categories, Keywords, Concepts & Categories Searching, Topics/Related Topics Navigation, Navigate Alphabetically, Enter Queries, Others
12	Product Platforms	Window NT/2000, Linux/Unix, Sun Solaris system, others

The increasing volume of electronic records coupled with the frequency of records changes require the development and implementation of taxonomies to maximize efficient retrieval of records [30]. Previous study by [31] classify CBIS into three distinct elements: information support; decision support; and communication support. [28] combines numerous approaches from research and practice on taxonomies, classification and ontologies, to develop a knowledge-cycle driven framework for understanding and then developing corporate taxonomies for effective exploitation of an organization's valuable knowledge resources.

As difference scenario projected different taxonomy, most of the time, produced taxonomies only matched the studied environment and requirements. Since there is no knowledge taxonomy for OMS in IHL context produced before, a specific and comprehensive study is necessary. This study aims to understand what are the base knowledge components to produce the intended and specific OMS taxonomy (IHL Malaysia). The focus would be on individual knowledge located in the technologies in the IHL organization. Developing our own taxonomy, in our own environment is an advantage because we will not be influenced on the classification of item that is not necessary or not even existed in our organization.

3 Methodology

The objective of the study is to propose a k-taxonomy for OMS in the context of IHL. Analyses of literature related to the OMS are discussed in the previous section become the basis on the OMS development. In the context of this research, individual knowledge is located as the initial stage of knowledge. Observations into the archives of public IHL reveals that knowledge resources can be categorized into three main sources namely paper documents, computer documents and self memory. This is in line with OMIS proposed in 1997 (Table 3). The three components of OMS are mapped into the six OM or CM types proposed by [14] which are non-computational, documents bases, knowledge based, case based, construction of distributed and combination of several techniques OM or CM. The knowledge types have been aligned to knowledge management technology available in IHL. The findings of the analysis are fitted in the taxonomy developed and discussed in the next section.

4 Result and Discussion

Proposed taxonomy is as follows:

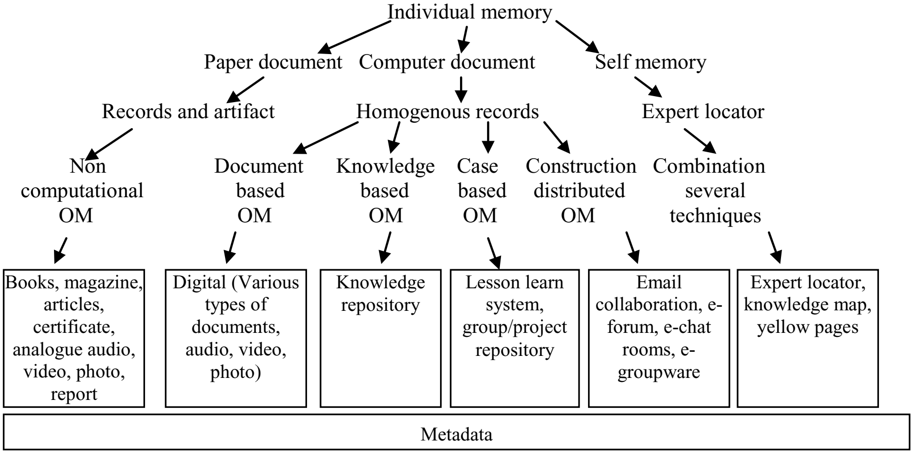

Individual memory
Paper document
Computer document
Self memory
Records and artifact
Homogenous records
Expert locator
Non computational OM
Document based OM
Knowledge based OM
Case based OM
Construction distributed OM
Combination several techniques
Books, magazine, articles, certificate, analogue audio, video, photo, report
Digital (Various types of documents, audio, video, photo)
Knowledge repository
Lesson learn system, group/project repository
Email collaboration, e-forum, e-chat rooms, e-groupware
Expert locator, knowledge map, yellow pages
Metadata

Fig. 1. Proposed OMS taxonomy

Metadata is the bottom layer of the proposed IHL OMS taxonomy. It plays an important role because it is the gist of knowledge and simplifies information about knowledge that leads to better identification of contents. Metadata describe knowledge kept and should be share among chosen components in the OMS taxonomy. It is also could be put as part of search engine components driving specific knowledge categorization. This will form values of knowledge worth transferring in OMS. Use of metadata helps to identify specific knowledge to be posted into repository. It is crucial that expert identify what is the related content to be highlighted before posting their knowledge.

[32] points that we must learn to capture the decision, the rationale behind the decision, the open questions related to the decision, the assumptions behind the decision, and any related supporting information. This would be important to describe the scenario or circumstances at that time supporting the existing of the knowledge. Another aspect that should be included in metadata is a very short conclusion of the story that relates to people's behavior or action that is moral. All good stories should end with a *moral* [33]. Moral reasoning involves how ethical decisions are arrived at, i.e., on what basis these decisions are supported or justified [34].At this level, the attributes describe the knowledge itself as well as its applicability in a context. Moral should cover the preventive action necessary to avoid problematic scenario.

Attributes that describe the knowledge itself include *aboutness, gist, ontological mappings and Web Ontology Language specifications* [35]. [28] notes standardized descriptive metadata with networked objects has the potential for substantially improving resource discovery capabilities by enabling field-based (e.g. *author, title, abstract, keywords*) searches, permitting indexing of non-textual objects, and allowing access to the surrogate content that is distinct from access to the content of the resource itself. Besides using the suggested metadata, this study identified its metadata for OMS through previous OMS studies in table 1. Finally it is concluded that metadata for OMS taxonomy are; *author, title, abstract (aboutness), keywords, subject category (gist ontology mapping), rational of decision/solution, moral/lesson learned, access (knowledge/expert), impact (individual, group, organization), type of story (success/failure)knowledge resources and knowledge physical location.*

The taxonomy and metadata found in this study will be the main element in designing OMS in IHL. Using KM technology as the base of components should interlink the existing tools in organization with organizational memories especially in providing cross functional knowledge sharing to organization. Since unit analysis of this study is individual academicians, individual memory is the uppermost layer in the taxonomy. Knowledge resources as detailed in Figure 1. will function as knowledge feeder to the OMS. In proposed OMS framework, all knowledge resources will go through OMS interface. This is where metadata will be extracted from those knowledge resources. Extracted metadata will be located into OMS central repository. OMS repository will consist all academician related activities and contributions. IHL able to use this repository to locate previous and existing knowledge in the organization. The repository able to produce academician's key performance indicator report extracted from all the knowledge resources in the organization. Figure 2 depicts OMS framework base from individual system perspective.

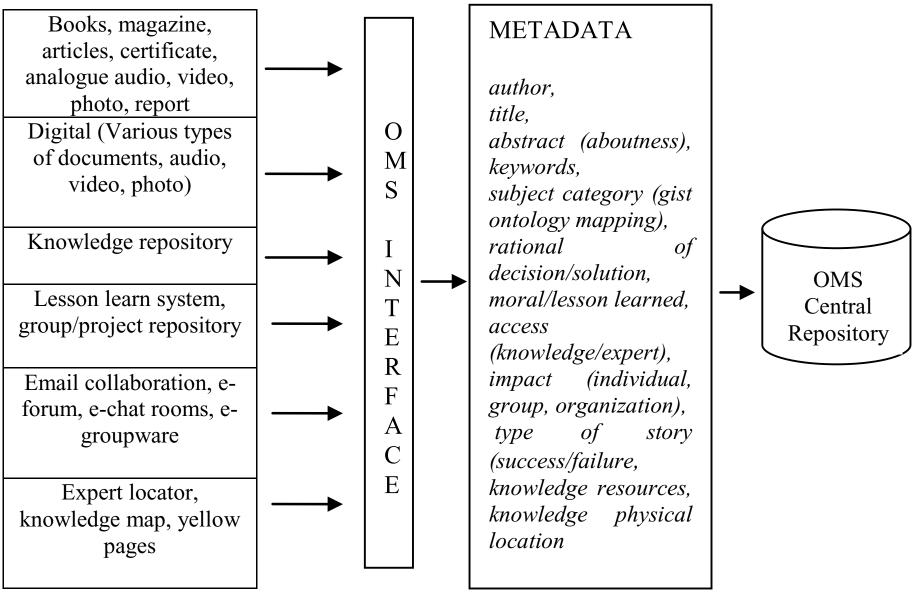

Books, magazine, articles, certificate, analogue audio, video, photo, report
Digital (Various types of documents, audio, video, photo)
Knowledge repository
Lesson learn system, group/project repository
Email collaboration, e-forum, e-chat rooms, e-groupware
Expert locator, knowledge map, yellow pages
OMS INTERFACE
METADATA
author, title, abstract (aboutness), keywords, subject category (gist ontology mapping), rational of decision/solution, moral/lesson learned, access (knowledge/expert), impact (individual, group, organization), type of story (success/failure, knowledge resources, knowledge physical location
OMS Central Repository

Fig. 2. OMS framework from individual system perspective

5 Conclusions and Future Work

This research reviews pertinent literature on OM computer base in the effort to develop OMS taxonomy for IHL. From the literature review researcher found that there are common understandings about forms and categorization of knowledge in organization. Observations are done in selected IHL and the findings from literature review are mapped into the synthesis of OMS components for a practical result. The OMS taxonomy and metadata produced filled up OMS framework from individual system perspective. Further study can be conducted to validate and verify the developed taxonomy.

References

1. Ghosh, R., Suete, L.: Knowledge Sharing: a Global Challenge. In: Arrow-Nelson Workshop, Turin (2006)
2. Vasconselos, J.B., Kimble, C., Rocha, A.: Ontologies and the Dynamics of Organizational Environment: An Example of Group Memory System for the Management of Group Competencies. In: Proceedings of I-KNOW, Graz (2003)
3. Md. Daud, R. A.: The K-Mapping and KM Implementation Strategy in Organization: A Perspectives. Sigma Rectrix Systems (M) Sdn Bhd (2005)
4. Klein, J.H.: Some Directions for Research in Knowledge Sharing. Knowledge Management Research and Practice, 6(1), 41–46 (2008)

5. Norris, D.M., Mason, J., Robson, R., Lefrere, P., Collier, G.: A Revolution in Knowledge Sharing. EDUCAUSE Review, 15–26 (2003)
6. Awad, E.M., Ghaziri, H.M.: Knowledge Management. Pearson Education Inc., London (2004)
7. Wasileski, J.S.: Learning Organization Principles & Project Management. In: ACM SIGUCCS 2005, pp. 435–439 (2005)
8. Nonaka, I., Konno, N., Toyama, R.: Emergence of 'Ba'. A Conceptual Framework for the Continues and Self-transcending Process of Knowledge Creation in Knowledge Emergence. In: Social, Technical and Evolutionary Dimensions of Knowledge Creation, pp. 13–29. Oxford University Press, Oxford (2001)
9. Irani, Z., Amir, M.S., Love, P.: Mapping Knowledge Management and Organizational Learning in Support of Organizational Memory. International J. Production Economics, 200–215 (2009)
10. Bencsik, A., Lıre, V., Marosi, I.: From Individual Memory to Organizational Memory (Intelligence of Organizations), World Academy of Science, Engineering and Technology (2009)
11. Hamidi, S.R., Jusoff, K.: The Characteristic and Success Factors of an Organizational Memory Information System. Computer and Information Science J. 2(1), 142–151 (2009)
12. Lehner, F., Maier, R.: How can organizational Memory Theories Contribute to Organizational Memory System? Information System Frontiers 2(3), 277–298 (2000)
13. Guerrero, L.A., Pino, J.A.: Understanding Organizational Memory. In: Proceedings of XXI International Conference of Chilean Computer Science Society, SCCC 2001, pp. 124–132. IEEE CS Press, Punta Arenas (2001)
14. Dieng, R., et al.: Methods and Tools for Corporate Knowledge Management, `http://ksi.cpsc.ucalgary.ca/KAW/KAW98/dieng`
15. Abdul Rahman, A., Hamidi, S.R.: Organizational Memory Information System Case Study in Faculty of Computer Science and Information System UTM. In: International Conference on Technology Management, Putrajaya (2006)
16. Morrissey, S.: The Design and Implementation of Effective Knowledge Management Systems. unpublished thesis (2005)
17. Weber, R.O.: Addressing Failure Factors in Knowledge Management. The Electronic J. of Knowledge Management 5(3), 257–347 (2007)
18. Mahesh, K., Suresh, J.K.: What is the K in KM Technology. The Electronic J. of Knowledge Management 2(2), 11–22 (2004)
19. Sangani, P.: Knowledge Management Implementation: Holistic Framework Based on Indian Study, `http://aisel.aisnet.org/pacis2009/69`
20. Krogh, V.G.: Individualist and Collectivist Perspectives on Knowledge in Organizations: Implications for Information Systems Research. J. of Strategic Information Systems, 119–129 (2009)
21. Butler, T., Feller, J., Pope, A., Emerson, B., Murphy, C.: Designing a core IT artifact for Knowledge Management System Using Participatory Action Research in a Government and a Non-Government Organization. J. of Strategic Information Systems 17(4), 249–267 (2008)
22. Mohayiddin, M.G., et al.: The Application of Knowledge Management in Enhancing the Performance of Malaysian Universities. The Electronic J. of Knowledge Management 5(3), 301–312 (2007)
23. Muhammad, S., Nawi, H.S.A., Lehat, M.L.: Taxonomy of K-Portal for Institute of Higher Learning in Malaysia: a Discovery. In: International Conference of Information Retrieval & Knowledge Management, Shah Alam, pp. 358–361 (2010)

24. Haron, H., Noordin, S.A., Alias, R.A.: An Interpretive Exploration on Tacit Knowledge Dimensions in Academia. In: Proceedings International Conference in Information Retrieval and Knowledge Management, pp. 325–330 (2010)
25. Riihimaa, J.: Taxonomy of Information and Telecommunication Technology System Innovations Adopted by Small and Medium Enterprises, Department of Computer Sciences, University of Tempere, Finland, Unpublished thesis (2004)
26. Nickerson, R.C., Varshney, U., Muntermann, J., Isaac, H.: Taxonomy Development in Information Systems: Developing a Taxonomy of Mobile Applications. In: 17th European Conference on Information Systems, pp. 1–13 (2009)
27. Woods, E.: Building a Corporate Taxonomy: Benefits and Challenges. Ovum (2004)
28. Sharma, R. S., Chia, M., Choo, V. and Samuel, E.: Using A Taxonomy for Knowledge Audits : Some Fields Experiences, `http://www.tlainc.com/articl214.htm`
29. Barquin, R.C.: What is Knowledge Management? Knowledge and Innovation. J. Knowledge Management Consortium International 1(2), 127–143 (2001)
30. Cisco, S. L, Jackson, W. K.: Creating Order out of Chaos with Taxonomies. Information Management J., `http://findarticles.com/p/articles/mi_qa3937/is_200505/ai_n13638950/`
31. Gilchrist, A.: Corporate Taxonomies: Report on a Survey of Current Practice. Online Information Review 25(2), 94–102 (2001)
32. Mentzas, G.: A Functional Taxonomy of Computer Based Information Systems. International J. of Information Management 14(6), 397–410 (1994)
33. Conklin, J.: Designing Organizational Memory: Preserving Intellectual Assets in a Knowledge Economy CogNexus Institute, unpublished thesis (2001)
34. Girard, J.P., Lambert, S.: The Story of Knowledge: Writing Stories that Guide Organizations into the Future. The Electronic J. Knowledge Management 5(2), 161–172 (2007)
35. Johnson, D.G.: Computer Systems: Moral Entities But Not Moral Agents. Ethics and Information Technology 8(1), 195–200 (2006)

A Framework for Optimizing Malware Classification by Using Genetic Algorithm

Mohd Najwadi Yusoff and Aman Jantan

School of Computer Science,
Universiti Sains Malaysia,
Penang, Malaysia
mohd.najwadi@gmail.com, aman@cs.usm.my

Abstract. Malware classification is a vital in combating the malware. Malware classification system is important and work together with malware identification to prepare the right and effective antidote for malware. Current techniques in malware classification do not give a good classification result when it deals with the new and unique types of malware. For this reason, we proposed the usage of Genetic Algorithm to optimize the malware classification system as well as help in malware prediction. The new malware classification system is based on malware target and its operation behavior. The result from this study will create a new framework that designed to optimize the classification of malware. This new malware classification system also has an ability to train and learn by itself, so that it can predict the current and upcoming trend of malware attack.

Keywords: Malware Classification, Genetic Algorithm, Unique Malware.

1 Introduction

Malware classification is one of the main systems in malware detection mechanism. It is used to classify the malware into its designated classes. Malware classification system that used machine learning techniques for classifying task was commercially applied in many anti-malware products such as Avira, AVG, Kaspersky, McAfee, Trend Micro, Symantec, Sophos and ESET [1]. Nowadays, malware commonly classified as Virus, Worms, Trojan Horses, Logical Bombs, Backdoor, Spyware, Exploit and Rootkit [2-3].

Malware classification system is very necessary and highly important when combating the malware [4]. In malware classification system, the main appliance or engine is named as classifier and it is used to classify the malware into the appropriate malware class. According to Filiol, current malware classes are mainly based on malware specific objective [3]. As for an example, malware in worm's classes used a network to send the copies of itself to other computer just to spread and do not attempt to alter the system. Therefore, by looking at the malware class, the right and effective antidote can be produced from malware specific objectives and this will help anti-malware products to prevent the malware from affecting the system.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 58–72, 2011.

In the recent advent, the widespread of malware has increased dramatically due to the fact that malware writers started to deploy an avoidance technique to avoid detection and analysis by anti-malware products [5]. By using this technique, malware writers can change the malware syntax or also known as malware signature but not its intended behavior, which has to be preserved [6]. The common avoidance technique used by malware writers is the code obfuscation. There are two types of code obfuscation which are polymorphism and metamorphism technique. Many new variants of polymorphic and metamorphic malware can easily be created by encrypting the code, flow transposition, substitution and renaming variable [7]. The other techniques to avoid detection are packing, anti-debugging and anti-virtualization.

The design of malware and its intention has become more sophisticated and significantly improved [8]. Current machine learning classifiers in malware classification system do not give a good classification result when it deals with the new and unique types of malware. It is because malware are becoming increasingly specialized and difficult to analyze [9]. New and unique types of malware are no longer can be classify easily to the current malware classes. These variants of malware have numerous attributes, combination syntax but showing the same behavior [9-10]. Thus, classification of malware has become more complicated and a new classification system is urgently needed.

In this paper, we proposed a framework that optimizes the current machine learning classifier by using Genetic Algorithm (GA). GA is a heuristic search that simulates the process of natural evolution. According to Mehdi, this heuristic algorithm is regularly used to generate useful solutions in optimization and search problems [11]. Malware that was created by an avoidance technique, providing almost the same functionality and showing the same behavior can be classify by using GA. It is because these types of malware basically worked same like crossover and permutation operation in GA [12]. As stated by Edge, GA has an ability to be a learning algorithm. As a result, it will make the new classification system become more reliable than the current classification systems [13].

The rest of the paper is organized as follows. Section 2 provided information about malware and its background. Section 3 is devoted to discuss about malware class and techniques involved. Proposed framework is presented in the section 4. Finally, Section 5 gives a few remarks as the conclusion.

2 Malware Background

2.1 Motivation

In this new millennium era, malicious code or malware has been recognized as the major threats to the computing world [10], [14]. All malware have their specific objective and target, but the main purpose is to create threats to the data network and computer operation [15]. Malware highly utilized the communication tools to spread itself. For examples, worms are being sent out through email and instant messages, Trojan horses attacked from infected web sites and viruses is downloaded from peer-to-peer connections to the user systems. According to Filiol, malware will pursue to abuse existing vulnerabilities on the systems to make their entry silent and easy [3]. In addition, malware works to remain unnoticed, either by actively hiding or simply not making its presence on a system recognized to the user.

With the development of the underground economy, malware is becoming very profitable product as the malware writer used to spam, steal information, perform web frauds, and many other criminal tasks. According to Martignoni, Malware had established during the past decade to become a major industry [9]. The new malware keep on increasing from time to time and this delinquent is seriously concerned by the security group around the world. Panda Security Lab reported that one third of existing computer malwares were created between Jan-Oct 2010 [16].

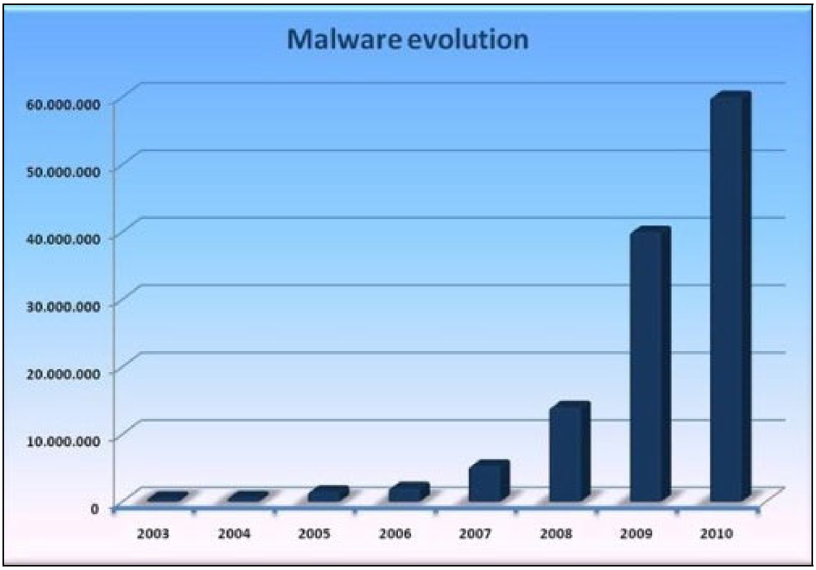


Fig. 1. Malware Evolution: Panda Lab Security [16]

The exponential growth of malware is also reported by many other security groups such as F-Secure, McAfee, Kaspersky, Sophos, Symantec and G-Data [17-22]. The increasing of malware has causes a billion of loss to the computer operation worldwide by breaking down the computer, congest the network, fully utilize the processing power, and many more bad impacts. According to ESET, more than half numbers of the malware samples nowadays are classified as unique. This unique malware are created by assimilating the existing malware with an avoidance technique. Fig. 2 is shows the total number of unique malware samples reported by ESET in 2005-2010.

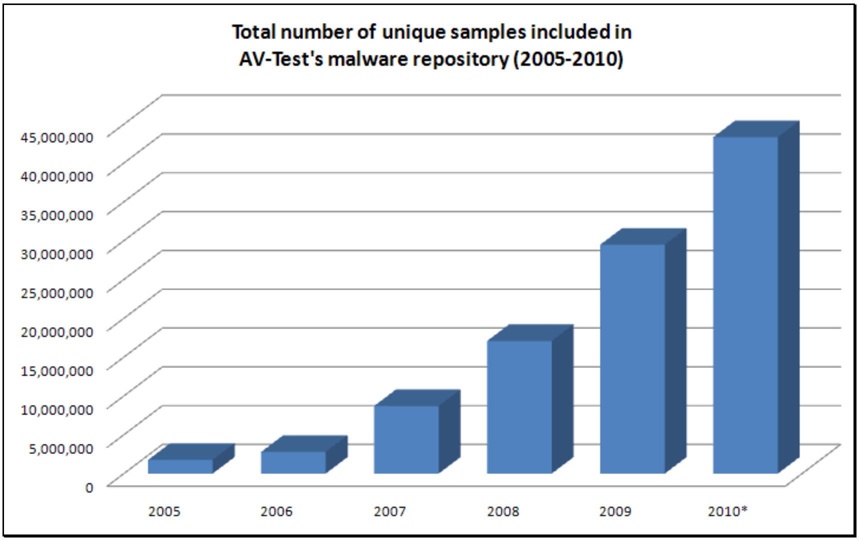


Fig. 2. Unique Malware Sample: ESET Threat Center [23]

2.2 Malware Avoidance Techniques

At the present time, nearly all modern malware has been implemented with a variety of avoidance techniques in order to avoid detection and analysis by anti-malware product [5]. The avoidance techniques that practically used by malware writers are code obfuscation, packing, anti-debugging and anti-virtualization. Code obfuscation technique can be divided into two types which are polymorphism and metamorphism. All these techniques can change the malware syntax but not its intended behavior, which has to be preserved [6]. The main purposed of these techniques is to avoid detection and to make the analysis process became more complicated [8].

A polymorphic technique can change the malware binary representation as part of the replication process. This technique consists of encrypted malicious code along with the decryption module [24]. It also has the polymorphic engine to decrypt and generate new mutants for the code before running it. When the polymorphic malware infecting the computer systems, it will encrypt itself by using new encryption key and a new code is generated. It is also has the polymorphic engine to decrypt and generate new mutants for the code before running it.

As for metamorphic technique, malware will transform the representation of programs by changing the code into different ways when it replicates but it still performs the same behaviors [25]. This technique can include control flow transposition, substitution of equivalent instructions and variable renaming [7]. Metamorphic malware can reproduced itself into different ways to rapidly created new malware variants and never look like the old malware.

Malware writers used packing technique to compress the Win32 portable execution file (PE file) and the actual file is unpacked as it is executed [26]. The main purposed of this technique is to protect the commercial software code from crack. A packed program contains with a program that is used for decompressing the compressed data during execution in the objective of making the task of static analysis become more difficult [27]. Packers will compress and sometimes it will encrypt the program. However, the program is transparently decompressed and decrypted at runtime when the program is loaded into memory. Some malware even packed its code several times in order to make it harder to be unpacked and used up so many resources until the computer hang or crash.

Debugger is a useful tool for reverse engineering of the malware code. A debugger normally can step through each instruction in a program code before it is executed. This tool performs their monitoring by either inserting breakpoints in the program or by tracing the execution using a special system calls [28]. Malware writers applied an anti-debugging technique to detect and avoid their malware from run under a debugger. Anti-debugging is an active technique where the malware writers embed code aimed to check process list for debugger process [29]. This technique will change the functionalities of the program when it interpreted by a debugger and make that program to discontinue it malicious intent or jump to end it.

The other technique is an anti-virtualization. Virtual environment is a place commonly used to do an analysis and extract the features of the malware. To avoid the malware from being analyze, malware writers used anti-virtualization to create malware code that has a capability to detect virtualization environment [30]. By using this technique, malware can check whether they are running in a virtual or normal environment before the execution. When the virtual environment is detected, malware might simply act like a genuine program or commonly refuse to run inside the virtual environment [31].

As an effect, it is not surprising that malware writer often use all these technique to automatically avoid detection. Even if anti-malware can detect it, that malware is unique and does not represent any existing malware class. Classification of malware become much harder and a new classification system is urgently needed.

3 Classification Part

3.1 Malware Class

Nowadays, malware are classified based on malware specific objective. Malware can be divided into several classes and there are diversities among the researchers in classifying the malware. In 2006, Aycock defined that malware consist of ten classes which are Logic Bomb, Trojan Horse, Back Door, Virus, Worm, Rabbit, Spyware, Adware, Hybrids and Zombies [2]. In 2009, Apel reported that malware consist of seventeen classes, seven more classes from Aycock [5]. Recent study by Filiol in 2010 stated that malware classes are divided into two groups which are “Simple” and “Self-Reproducing” [3]. However, the common class of malware nowadays is likely represented in Table 1.

Table 1. Common malware classes

Types	Specific Objective
Virus	Computer program that can copy itself and infect a computer.
Worms	Self-replicating malware which uses a computer network to send copies of itself to other nodes without any user intervention.
Trojan Horse	Software that appears to perform a desirable function for the user prior to run or install, but harms the system
Logical Bombs	Simple type of malware which wait for significant event such as date or action to be activated and launch its criminal activity
Backdoor	Method of bypassing normal authentication, securing remote access to a computer, while attempting to remain undetected.
Spyware	Malware that can be installed on computers, and collects small pieces of information about users without their knowledge
Exploit	A piece of software that attacks particular security vulnerability. Exploits are not necessarily malicious in intent
Rootkit	Software inserted onto a computer system after an attacker has gained control of the system.

We did an analysis about the malware operation behavior. For this analysis, we focus on worms and Trojan horses only. It is because these two are the common types of malware that attack host-based system [12]. We are observing the executed malware by focusing it into the specific target and its operation behavior in windows environment systems. The malware samples that we are using for this experiment comprises of 300 unique samples and the result that we get is shows in Fig. 3.

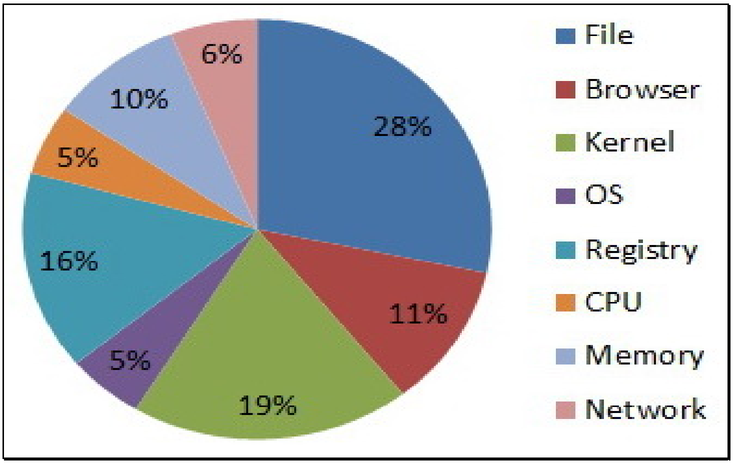


Fig. 3. Malware Specific Operation

From this result, we can classify the malware specific operation into 4 main classes which are Data, Application, System and Dos. Malware that are attack File will group under Data class and malware that are attack Browser will be under Application class. Malware that are attack Kernel, Operating System and Registry will be group under System class. Lastly, malware that are attack CPU, Memory and Network will be group under Dos class.

3.2 Machine Learning Classifier

Classifier is the main engine in the malware classification system. In this paper, we had studied several current machines learning classifier that has been used in malware classification. Table 2 shows the summary of Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and K-Nearest Neighbor (KNN).

Table 2. Machine learning classifier in malware classification

Classifier	Speed	Accuracy	Strength	Weakness
Naïve Bayes [32-33]	Very Fast	High	Fast, easier to maintain and consistence result	Sensitive to the correlated attributes
Support Vector Machine (SVM) [4], [34]	Fast	High	Regression and density estimation results. Better performance in text classification, pattern segment and spam classification	Expensive and problem lies on the prohibitive training time.
Decision Tree [33], [35]	Very Fast	High	Easy to understand, easy to generate rules and reduce problem complexity	Mistake on higher level will cause all wrong result in sub tree
K-Nearest Neighbor [36]	Slow	Moderate	Useful when the dependent variable takes more than two values and effective if the training data is large	Very computationally intensive. $O(n^2)$

Based on the summary above, we decided to select DT classifier to be used and work together with the GA in the proposed framework. DT is selected because this algorithm is the most suitable algorithm for our proposed system that is based on malware specific target and its operation behavior. Malware target is referring to the Windows file system which is in the tree format. As for example, *TrojanDropper:Win32* malware has targeted two directories to execute and perform its malware operation behavior [37];

1. *C:\Documents and Settings\Users\Local Settings\Temp\*
2. *C:\Windows\System32*

3.3 Genetic Algorithm

GA is a heuristic search that simulators the process of natural evolution. The standard GA can be seen as the combination of bit-string representation, with bit-exchange crossover and bit-flip mutation, roulette-wheel selection plus generational replacement. GA is belongs to the larger class of Evolutionary Algorithm (EA). GA include the survival of the fittest idea into a search algorithm which provides a method of searching which does not need to explore every possible solution in the feasible region to obtain a good result.

GA also commonly used on a learning approach to solve computational research problem. According to Mehdi, this heuristic algorithm is regularly used to generate useful solutions in optimization and search problems [11]. In a GA, a population of strings which encode candidate solutions to an optimization problem evolves toward better answers. By tradition, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible [38].

A simple presentation of GA is shows as follows;

```
generate initial population, G(0);
evaluate G(0);
t:=0;
repeat
      t:=t+1;
      generate G(t) using G(t-1);
      evaluate G(t);
until find a solution
```

GA technique is implemented in this research in order to solve and classify the unique malware that was created by an avoidance technique. A framework is proposed by combining GA with the current machine learning classifier. New and unique malware can be detected and classify through this technique. It is because, unique malware work similar like crossover and permutation operation in GA [12]. An avoidance techniques that normally used by malware writers can be detected and classified using this new malware classification system because it not only filter the content but also train and learn by itself, so it can predict the current and upcoming trend of malware attacks.

4 Proposed Framework

We proposed the usage of the GA to enhance and optimize performance of the new classifier. GA will work together with the DT and there will be the training phase for GA that is explained in the sub section. Our proposed framework consists of several elements which are Malware Sample, Virtual Environment, Knowledge Storage, Classifier and Class Operation. Fig. 4 shows our proposed framework for optimizing malware classification using GA. The next sub-section is explained about each element in this proposed framework.

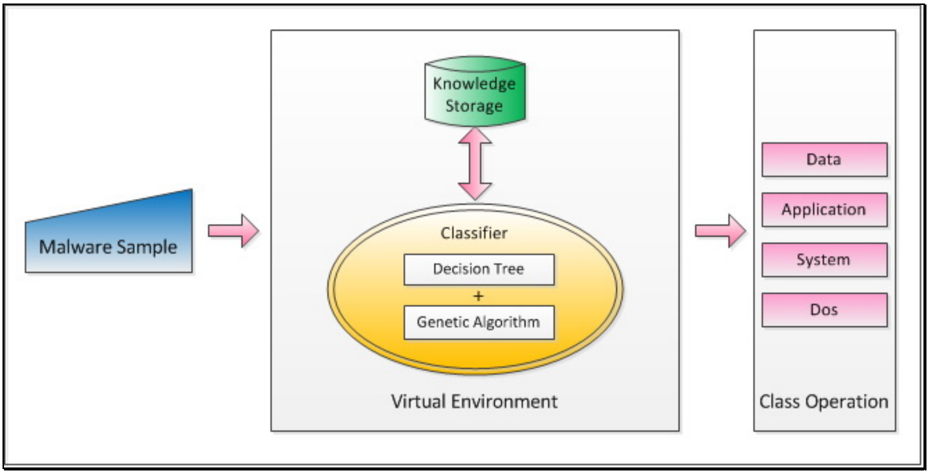


Fig. 4. A Framework for Optimizing Malware Classification Using Genetic Algorithm

4.1 Malware Sample

Most of malware are designed to attack windows-based operating system. According to Germany anti-virus firm G-Data (2010), 99.4% of the new malware are purposely target windows operating system [22]. Therefore, we are focusing our test on windows-based operating system only and narrowed down our scope and focus to Win32 portable execution file (PE file) that contains malicious code. This sample PE file is collected thru college network, internet and some suspicious execution file in windows operating system itself. Each of the file has been extracted to obtain their features and behavior. All malware targets and its operation behavior are compiled in malware profile and stored in the knowledge storage.

4.2 Virtual Environment

Virtual environment is a security mechanism for separating running programs. It is often used to execute untested code, or untrusted programs from unverified third-parties, suppliers and untrusted users. Virtual environment machine run isolated, so the benefit of a virtual environment machine is that it cannot directly modify the "real" operating system running on our machine. The virtual machine is assigned its own hard disk space, and that's what it treats as its virtual "entire hard disk".

In this proposed framework, we decided to use Windows XP Mode [39] in Windows 7 Platform as our testing platform. The purposed of this virtual environment is to create secure platform in order to minimize damage and attack to the actual machine when the malware sample is executed. Since virtual environment is assigned with its own hard disk space, if got infected, we can simply remove it from our test machine.

4.3 Knowledge Storage

Knowledge storage is the database storage where we stored all the malware profile after finish the analysis and features extracted. Malware profile consists of malware sample, MD5 hash, malware size, malware specific target and it class operation behavior. Table 3 shows an example of malware profile in our knowledge storage.

Table 3. Malware Profile in Knowledge Storage

Malware Sample	MD5 Hash	Size (byte)	Specific Target	Class Operation
umdmgr.exe	A2D09FE1EB487A799907527E494AA06B	61,440	• C:\Windows\System32	System
sdra64.exe	5A92608A111E80356F346E7405171C4A	127,488	• C:\Documents and Settings\[UserName]\Local Settings\Temp\ • C:\Windows\System32 • C:\Windows	Data
aa.exe	E9E1D7AD36C53C7B70C38DA1514BBD5D	58,749	• C:\Documents and Settings\[UserName]\Application Data • C:\Documents and Settings\All Users\Application Data • C:\Documents and Settings\All Users	Application
lsass.exe	09BA41F35106E94A9E5A3639AC52B280	22,528	• C:\Documents and Settings\[UserName]\Application Data • C:\Documents and Settings\All Users\Application Data • C:\Documents and Settings\All Users	Application

The knowledge storage is designed to be re-writable by the system. Certain unique malware has a relationship and link with the other malware when it is executed. This malware relationship is unable to analyze because during analysis process, malware is executed one by one [28-29]. At first, the information in the knowledge storage is not sufficient enough to be used in the classification system. For this reason, we used GA to conduct a training phase together with the DT. The system will update the database after gone thru the training phase.

4.4 Classifier

Classifier is the main engine in our malware classification system. We have selected the DT as our machine learning classifier and combine it with the GA. In this new system, we are having the classifier training phase and GA is used to become a learning algorithm. During the classifier training phase, we used different malware samples as the training data set. We must use different malware samples because we want to let the classifier to learn and update the new malware into malware profile. One of the main goals is to detect and classify the unique malware that has a relationship during the execution. The other goal is to find unique malware that perform the same behavior but providing different syntax representation. Fig.5 shows the classifier training phase process in our proposed malware classification system.

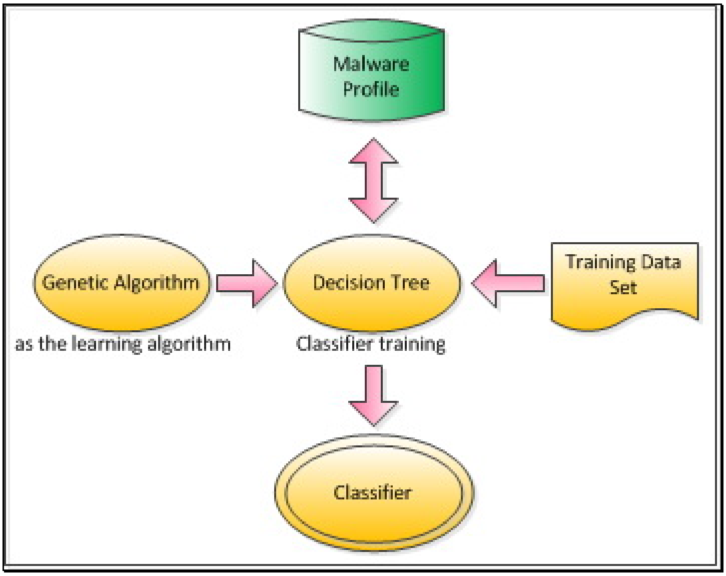


Fig. 5. Classifier Training Phase

As mention earlier, the malware profile in the knowledge storage is designed to be re-writable. Classifier will keep on updating the malware profile during this training phase. The classification result will become more accurate after this training phase. This process also shows the ability of GA in helping DT to optimize the classification process.

4.5 Class Operation

Class operation is our new proposed malware classes. Previously, we had done an analysis about the malware operation behavior using worms and Trojan horse. Based on that analysis, we proposed the malware class based on malware operation behavior as shows in Table 4.

The right and effective antidote is necessary in combating the malware. The antidote is produced by the other system in malware detection mechanism. Although the antidote is not produced by the classification system, the classification system can work together with other system in preparing the right antidote by looking at the malware class. The reason is, we need to find the cause and effect before come out with the solution which is antidote in our cases.

Table 4. Proposed malware class based on malware operation behavior

Class Operation	Rank	Attacked examples	Affected examples
Data	1	Malware attack office and adobe file	.doc, .ppt, .xls and .pdf file
Application	2	Malware attack application such as office application, audio application and video application	Microsoft Office, Winamp and Windows media Player
System	3	Malware attack the entire Operating System	Windows XP, Windows 7 and Windows Server 2008
Dos	4	Malware attack physical hardware and entire machine	CPU usage, memory and network access

In our proposed class, all four classes are related with each other in rank 1 to 4 starting with the Data class and end with the Dos class. If the classifier classified the malware sample in the Data class, the antidote is prepared based on Data operation and if the malware sample are classified in the Application class, the antidote is prepared based on Data and Application operation. It is same with the System and the Dos classes. Antidote is prepared for Data, Application and System if the malware sample is classified in System class and antidote for entire class is prepared if the malware sample is classified in the Dos class.

The reason we proposed malware class based on its operation is to reduce the process flow and cost in malware detection mechanism. For example, if the malware is in the Data class, the detection mechanism does not need to look and prepare the antidote for other classes because that malware only attack file without attempt to attack the application and operating system. If the malware attack Microsoft Office application which is under Application class, habitually it will also affect the office application file under Data class, but not necessary to attack that particular operating system, which is under System class.

However, not all malware will attack based on this rank class but it normally do so. Duplicating a file is considered under Data class but it also consumes and bottlenecks the memory usage, so for that cases, it is classified in the Dos class. The antidote is prepared for the Data and the Dos class only. All the decision of classification is made by DT and is trained by the GA to optimize the classification process.

5 Conclusion

In this paper, we have proposed a new framework for optimizing malware classification by using GA. By using this new way of classification system, new and unique types of malware can be classify by checking the similarity of malware target and its operation behavior. We also proposed a new malware classes which mainly based on malware operation behavior. This new classification system is important because current method of classification cannot detect and classify unique malware, hence drop down the performance of anti-malware products. There are several limitations in this research. We only focus on host-based machine and windows based operating system. The test subject is also limited to Worms and Trojan horses only. In order to improve efficiency and better classification performance, this research will continue with other domains and other malware types.

Acknowledgments. The author would like to thanks to the members in the Security Research Group, Universiti Sains Malaysia for their helpful discussion and suggestion. This work is supported by Short-term Grant No.304/PKOMP/639021.

References

1. Gheorghescu, M.: An Automated Virus Classification System. In: Virus Bulletin Conference Microsoft (2005)
2. Aycock, J.: Computer Virus and Malware. Springer, Heidelberg (2006)
3. Filiol, E.: Viruses and Malware. In: Handbook of Information and Communication Security, pp. 747–769. Springer, Heidelberg (2010)
4. Rieck, K., Holz, T., Willems, C., Düssel, P., Laskov, P.: Learning and Classification of Malware Behavior. In: Zamboni, D. (ed.) DIMVA 2008. LNCS, vol. 5137, pp. 108–125. Springer, Heidelberg (2008)
5. Apel, M., Bockermann, C., Meier, M.: Measuring Similarity of Malware Behavior. In: The 34th Annual IEEE Conference on Local Computer Networks, pp. 891–898. IEEE Press, Zurich (2009)
6. Preda, M., Christodorescu, M., Jha, S., Debray, S.: A Semantics-Based Approach to Malware Detection. Journal of Transactions on Programming Languages and Systems 30(5) (2007)
7. Szor, P.: The Art of Computer Virus Research and Defense. Symantec Press, Addison-Wesley Professional (2005)
8. Noreen, S., Murtaza, S., Shafiq, M., Farooq, M.: Evolvable Malware. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, pp. 1569–1576. ACM, New York (2009)
9. Martignoni, L., Paleari, R., Bruschi, D.: A Framework for Behavior-Based Malware Analysis in the Cloud. In: Prakash, A., Sen Gupta, I. (eds.) ICISS 2009. LNCS, vol. 5905, pp. 178–192. Springer, Heidelberg (2009)
10. Bayer, U., Habibi, I., Balzarotti, D., Kirda, E., Kruegel, C.: A View on Current Malware Behaviors. In: Proceedings of the 2nd USENIX Conference on Large-scale Exploits and Emergent Threats, USENIX, USA (2009)
11. Mehdi, S., Tanwani, A., Farooq, M.: IMAD: In-execution Malware Analysis and Detection. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation. ACM, New York (2009)
12. Zolkipli, M.F., Jantan, A.: Malware Behavior Analysis: Learning and Understanding Current Malware Threats. In: Second International Conference on Network Applications, Protocols and Services, pp. 218–221. IEEE Press, Kedah (2010)
13. Edge, K., Lamont, G., Raines, R.: A Retrovirus Inspired Algorithm for Virus Detection and Optimization. In: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, pp. 103–110. ACM, New York (2006)
14. Zhao, H., Xu, M., Zheng, N., Yao, J., Ho, Q.: Malicious Executable Classification Based on Behavioral Factor Analysis. In: Proceedings of the 2010 International Conference on e-Education, e-Business, e-Management and e-Learning, pp. 502–506. IEEE Press, Sanya (2010)

15. Zolkipli, M.F., Jantan, A.: A Framework for Malware Detection Using Combination Technique and Signature Generation. In: Proceedings of the Second International Conference on Computer Research and Development, pp. 196–199. IEEE Press, Kuala Lumpur (2010)
16. Panda Security Lab. One third of existing computer viruses were created in (January-October 2010) Panda, http://www.channeltimes.com/story/one-third-of-existing-computer-viruses-were-created-upto-october-2010-panda/
17. F-Secure IT Threats Security Summary, http://www.f-secure.com/en_EMEA-Labs/news-info/threat-summaries/
18. McAfee Labs, Malware Is Their Business...and Business Is Good, http://blogs.mcafee.com/mcafee-labs/malware-is-their-businessand-business-is-good
19. Kaspersky Security Bulletin. Malware Evolution (2010), http://www.securelist.com/en/analysis/204792161/Kaspersky_Security_Bulletin_Malware_Evolution_2010
20. Sophos – Security Threats Report: Mid-Year (2010), http://www.sophos.com/sophos/docs/eng/papers/sophos-security-threat-report-midyear-2010-wpna.pdf
21. Cyber War - Much Ado about Nothing or the Real Deal? http://www.invincea.com/blog/2010/07/cyber-war-much-ado-about-nothing-or-the-real-deal/
22. G-Data - Number of New Computer Viruses at Record High, http://www.gdatasoftware.co.uk/about-g-data/press-centre/news/news-details/article/1760-number-of-new-computer-viruses.html
23. ESET Threat Center, http://www.eset.com/us/threat-center
24. Vinod, P., Laxmi, V., Gaur, M.S.: Survey on Malware Detection Methods. In: Proceedings of the 3rd Hackers' Workshop on Computer and Internet Security, IITK, Kanpur, India (2009)
25. Zhang, Q., Reeves, D.: MetaAware: Identifying Metamorphic Malware. In: Twenty-Third Annual Computer Security Applications Conference, pp. 411–420. IEEE Press, Miami Beach (2007)
26. Han, S., Lee, K., Lee, S.: Packed PE File Detection for Malware Forensics. In: 2nd International Conference on Computer Science and its Applications, CSA, Korea (2009)
27. Alazab, M., Venkataraman, S., Watters, P.: Towards Understanding Malware Behavior by the Extraction of API Calls. In: Second Cybercrime and Trustworthy Computing Workshop, pp. 52–59. IEEE Press, Ballarat (2010)
28. Desfossez, J., Dieppedale, J., Girard, G.: Stealth Malware Analysis from Kernel Space With Kolumbo. Journal of Computer Virology 7(1), 83–93 (2011)
29. Liu, L., Chen, S.: Malyzer: Defeating Anti-detection for Application-Level Malware Analysis. In: Abdalla, M., Pointcheval, D., Fouque, P.-A., Vergnaud, D. (eds.) ACNS 2009. LNCS, vol. 5536, pp. 201–218. Springer, Heidelberg (2009)
30. Lau, B., Svajcer, V.: Measuring Virtual Machine Detection in Malware Using DSD Tracer. Journal of Computer Virology 6(3), 181–195 (2010)
31. Daewon, K., Ikkyun, K., Jintae, O., Jongsoo, J.: Behavior-Based Tracer to Monitor Malicious Features of Unknown Executable File. In: Fifth International Multi-Conference on Computing in the Global Information Technology, IEEE Press, Spain (2010)

32. Wang, C., Pang, J., Zhao, R., Fu, W., Liu, X.: Malware Detection Based on Suspicious Behavior Identification. In: First International Workshop on Education Technology and Computer Science, pp. 198–202. IEEE Press, Wuhan (2009)
33. Farid, D.M., Harbi, N., Rahman, M.Z.: Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection. International Journal of Network Security & Its Applications 2(2), 12–25 (2010); arXiv.org
34. Mezghani, D., Boujelbene, S., Ellouze, N.: Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification. International Journal of Hybrid Information Technology 3(3), 23–34 (2010)
35. Komashinskiy, D., Kotenko, I.: Malware Detection by Data Mining Techniques Based on Positionally Dependent Features. In: Proceedings of the 2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing, pp. 617–623. IEEE Press, USA (2010)
36. Hall, P., Park, B., Samworth, R.: Choice of Neighbor Order in Nearest-Neighbor Classification. Journal of the Institute of Mathematical Statistics 36(5), 2135–2152 (2008)
37. ThreatExpert - TrojanDropper:Win32, http://www.threatexpert.com/report.aspx?md5=045f8c12b349dafa8c0180a9237f5319
38. Cha, S.-H., Tappert, C.: A Genetic Algorithm for Constructing Compact Binary Decision Trees. Journal of Pattern Recognition Research 4(1), 1–13 (2009)
39. Windows XP Mode, http://www.microsoft.com/windows/windows-7/features/windows-xp-mode.aspx

Building a Corpus-Derived Gazetteer for Named Entity Recognition

Norshuhani Zamin and Alan Oxley

Department of Computer and Information Sciences
Universiti Teknologi PETRONAS,
31750 Tronoh, Perak, Malaysia

Abstract. Gazetteers, or entity dictionaries, are an important element for Named Entity Recognition. Named Entity Recognition is an essential component of Information Extraction. Gazetteers work as specialized dictionaries to support initial tagging. They provide quick entity identification thus creating richer document representation. However, the compilation of such gazetteers is sometimes mentioned as a stumbling block in Named Entity Recognition. Machine learning, both rule-based and look-up based approaches, are often used to perform this process. In this paper, a gazetteer developed from MUC-3 annotated data for the 'person named' entity type is presented. The process used has a small computational cost. We combine rule-based grammars and a simple filtering technique for automatically inducing the gazetteer. We conclude with experiments to compare the content of the gazetteer with the manually crafted one.

Keywords: Gazetteer, Named Entity Recognition, Natural Language Processing, Terrorism.

1 Introduction

Named Entity Recognition (NER) involves the identification of certain occurrences of words or expressions in unstructured texts and the classification of them into a set of predefined categories of interest. NER is often implemented as a pre-processing tool for an Information Extraction (IE) system. One application is document retrieval, or automatic document forwarding. For this application [1] states that "… documents annotated with NE information can be searched more accurately than raw text." As an example, NER annotation would allow the search for all texts mentioning the *company* "Hong Leong" such as the notable "Hong Leong Bank" and "Hong Leong Group", both Malaysian establishments. NER is supposed to ignore documents about unrelated companies of the same name. However, our investigations found that classification of certain entities often involves challenging ambiguities. Among the types of ambiguity is metonymy [2]. Metonymy is a figure of speech in which one word or phrase is substituted for another with which it is closely related; for instance, "England" is an *organization* in the statement "England won the world cup" while "England" is a *location* in the statement "The world cup took place in England". The use of syntactic features and word similarity is a possible solution for metonymy recognition, as described in [2].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 73–80, 2011.

NER systems are commonly classified into three categories: machine learning, rule-based and look-up based [3]. The machine learning approach requires an annotated training corpus to establish the learning process thus allowing it to predict the most likely entities in the given text. The machine learning approach has proved to be advantageous when there are either no terminologies or only partial ones. All of the basic machine learning techniques - supervised, semi-supervised and un-supervised - are only possible with the availability of huge amounts of training data. Nonetheless, the training data or corpus, which is often manually annotated, can be a really cumbersome task for a human to create.

Rule-based NER systems achieve the best results of the three categories in all NER evaluations. In the absence of learning/training data, rule-based NER shows promising results. However, the rules involved in rule-based NER are likely to be quite complex. This approach relies heavily on the knowledge of linguistic experts. Consequently, the hand crafted rules are difficult to maintain without help from the experts in supporting large-scale NER [4]. The look-up based approach is a straightforward method and provides an extremely efficient way to perform NER [5]. In general, the look-up based approach makes use of lists of common entities to provide clues. (The term "list" is often used interchangeably with the term "gazetteer", "lexicon" and "dictionary" [4].) It is a fast method with a small programming cost. The only processing required is to map the entities in the list against the given text. Additionally, lists are commonly acquired either from a corpus, the web or Wikipedia. However, NER systems based solely on such lists suffer from the limitations of coverage and ambiguity [6]; for instance, the word "Washington" in "President Washington" is clearly a *person* based on the external evidence as the regular expression "President" appears before the word. However, if the word "Washington" is found in the look-up list as a *location* (as in "Washington D.C."), then this entity will be identified as a location based on this internal evidence. This limitation shows that huge lists may also miss some important entities, but with proper techniques and methods the drawback is fixable.

In this paper, we investigate look-up based NER. The list to be looked up is referred to as a "gazetteer" throughout the paper. We propose a technique to automatically derive the gazetteer for name entities from a terrorism text corpus (MUC-3). Our experiment explored only the *person* entity type, by identifying the proper name, and the result is compared with the manually annotated list. We used a simple tokenizing technique for the entity extraction, with a dictionary filtering scheme inspired by [7], and combined it with simple lexical patterns [8] for the grammar rules.

2 Related Work

The task of automatically generating gazetteers for NER has been studied for many years. [9] uses lexical patterns to identify nouns from a similar semantic class. For instance, a noun phrase that follows "the President of" is usually the name of a country. The construction of such noun phrases are based on common patterns developed manually. These patterns are also referred to as grammar rules. Similarly, research in [10] uses lexical patterns but with a small number of entities called *seeds* to train the proposed bootstrapping algorithm in order to derive more related entities. Research in [5] shows that the use of a simple filtering technique for improving the automatically acquired gazetteer has contributed to a highly significant result, close to that of a state-of-the-art NER. A novel method for exploiting repetition of entities is presented in [7] with a highly efficient filtering technique to filter unwanted entities. The recall-enhancing approach requires an entire test set to be available despite substantially improving the extraction performance. Inspired by the work in [11], the researcher in [4] proposes an automatic approach to generate gazetteers based on initially defined entities *(seeds).* The bootstrapping algorithm in [10] is applied, with little modification, to handle novel types of named entities including car brands.

The use of the structural information of a language has recently been studied as one of the permissible approaches to automatically induce gazetteers from texts. The research in [12] shows a successful attempt at generating gazetteers from a Japanese corpus using the cache features, coreference relations, syntactic features and case-frame features of the language. This information is commonly available from structural analysis done by linguists. It has been observed from the evaluation results that the use of language specific features improves the performance of the system. Meanwhile, a novel method using the significant high-frequency strings of the corpus is introduced in [13]. The method uses the distribution of these strings in the document as candidate entities to filter the invalid entities. Additionally, the research team extends the work by incorporating word-level features and has successfully induced a gazetteer from Chinese news articles at around 80% accuracy.

More recently, Natural Language Processing research in [14] uses concepts and instance information from ontologies for NER and IE systems. The research automatically generates gazetteers from a corpus using the `rdf:type` information. An RDF stylesheet is defined and used to select statements about instances of relevant concepts. These instances are then converted to structured gazetteer source files. To the best of our knowledge, none of the discussed research generates gazetteers from the MUC-3 text corpus.

3 Corpus

A text corpus is a collection of text. Most corpora are designed to contain a careful balance of material in one or more genres. Commonly, information in corpora is unstructured. There is a wide range of corpora available, such as the collections of speeches[1], e-books[2], newswire articles[3] and texts of multiple genres. The Brown Corpus[4], which was established in 1961, is the pioneer corpus. It is a collection of 500 English text sources categorized by different genres. Some of these corpora contain linguistic annotations, representing part-of-speech tags, named entities, syntactic

[1] http://www.tlab.it/en/allegati/esempi/ inaugural.htm

[2] http://www.gutenberg.org

[3] http://www.reuters.com

[4] http://mailman.uib.no/public/corpora/2005-June/001262.html

structures, semantic roles, etc. However, most of the annotated corpora are not publicly accessible, such as the British National Corpus[5].

In this research we work with a terrorism corpus to support our ongoing research, which aims to develop a counterterrorism IE mechanism [15-17]. The series of Message Understanding Conferences (MUCs) funded by the Defense Advanced Research Projects Agency (DARPA) has established seven types of corpora, two of which are collections of American terrorism texts. The goal of these conferences was to encourage the development of new and better methods of IE. We use the MUC-3 corpus, a collection of news records on terrorism from Latin America. Unfortunately, this corpus is unannotated. The pre-processing part of our gazetteer generation process relies on the part-of-speech tags of the words to identify a possible group for an entity. A free part-of-speech tagger is recycled to perform the tagging process.

4 Gazetteer Generation

The framework of our automatic corpus-based gazetteer generation process is illustrated in Fig. 1. In the first step, to avoid starting from scratch, we adopted a free part-of-speech (POS) tagger known as Brill's Tagger [18] to assign possible POS tags to the words in the MUC-3 text corpus. Next, each word and its corresponding POS tag is tokenized by the *Word/Tag Tokenizer* module. The *Entity Extractor* module extracts all the words that have been assigned a 'proper noun singular' or 'proper noun plural' tag, as represented by NNP and NNPS, respectively, in the Brill's Tagger notation. Additionally, grammar rules are applied to potentially disambiguate the problem discussed earlier. The grammar rules adopted in [8] are the common lexical patterns or the regular expressions found in English text, i.e. the context around the proper names that identifies their type. Following are several regular expressions used to identify a person's name:

1. @Honorific CapitalizedWord CapitalizedWord
   a. @Honorific is a list of honorific titles such as General, Lieutenant, Captain, etc.
   b. Example: General Ramon Niebels
2. CapitalizedWord CapitalLetter @PersonVerbs
   a. @PersonVerbs is a list of common verbs that are strongly associated with people such as *{*said, met, walked, etc.*}*
3. @FirstNames CapitalizedWord
   a. @FirstNames is a list of common first names collected from the corpus.
   b. Example: Maza Marquez
4. CapitalizedWord CapitalizedWord [,] @PersonSuffix
   a. @PersonSuffix is a list of common suffixes such as *{*Jr., Sr., II, III, etc.*}*
   b. Example: Mark Green, Jr.
5. CapitalizedWord CapitalLetter [.] CapitalizedWord
   a. CapitalLetter followed by an optional period is a middle initial of a person and a strong indicator that this is a person name.
   b. Example: President Peace R.Pardo

[5] `http://www.natcorp.ox.ac.uk`

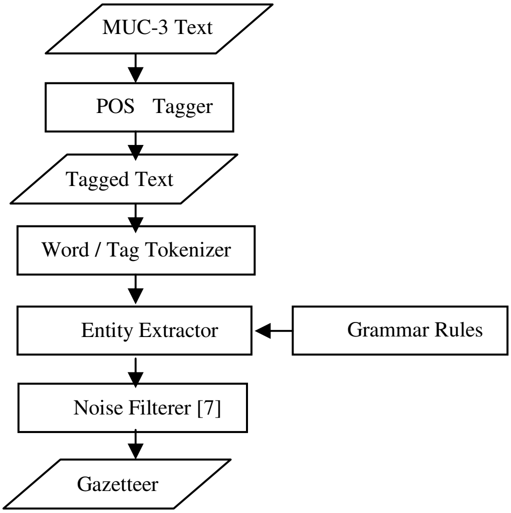


Fig. 1. Gazetteer Generation Framework

A dictionary matching scheme is often vulnerable to false positives. A false positive is a case where some proper names identified by the entity extractor are in fact non-names and can be considered as noise. False positives often degrade such a system's accuracy. Hence, we added the *Noise Filterer* module to the framework to remove the unwanted names by simply eliminating low-confidence predictions.

There are two metrics used in this module as introduced in [7]: Predicted Frequency (PF); Inverse Document Frequency (IDF). The PF metric estimates the degree to which a word appears to be used consistently as a name throughout the corpus.

$$PF(w) = \frac{cpf(w)}{ctf(w)} \tag{1}$$

Here, *cpf(w)* is the number of times that a word *w* is identified as a name and *ctf(w)* is the number of times it appears in the entire test corpus. Inverse Document Frequency is calculated using the IDF metric.

$$IDF(w) = \frac{\log(\frac{N+0.5}{df(w)})}{\log(N+1)} \tag{2}$$

Here, *df(w)* is the number of documents that contain the word *w* and *N* is the total number of documents in the corpus. *IDF* is a suitable metric for person name recognition since this type of entity does not frequently appear as an English word. Both metrics return a result between 0 and 1. A measure which combines these two metrics multiplicatively, giving a single probability of a word being a name and how common it is in the entire corpus, is as follows:

$$PF.IDF(w) = PF(w)xIDF(w) \tag{3}$$

A word with low *PF.IDF* score is considered ambiguous in the corpus and is excluded from the gazetteer.

5 Experimental Results

The system was evaluated using the *Precision (P)* and *Recall (R)* metrics. Briefly, *Precision* is the proportion of names proposed by the system which are true names while *Recall* is the proportion of true names which are actually identified. These metrics are often combined and referred to as the *F-Measure (F)*. Hence, the *F-Measure* is a weighted harmonic between *P* and *R*.

$$Precision\ (P) = correct / (correct + wrong) \tag{4}$$
$$Recall\ (R) = correct / (correct + missed) \tag{5}$$
$$F\text{-}Measure\ (F) = 2PR / (P + R) \tag{6}$$

Here, *correct* is the number of names extracted by the system that are persons names, *wrong* is the number of names extracted by the system that are not persons names while *missed* is the number of persons names that are extracted manually but not by the system. Our experiment was conducted on 40 randomly selected texts from the MUC-3 corpus. The same set of data is used on the Stanford NER [19]. The Stanford Named Entity recognizer uses Conditional Random Field (CRF) method and was trained on Conference on Computational Natural Language Learning (CoNLL), MUC-6, MUC-7 and Automatic Content Extraction (ACE) named entity corpora with a fairly results across domains. CRF is a type of discriminative undirected probabilistic graphical model which each vertex represents a random variable whose distribution is to be inferred. Edges correspond to dependencies between two random variables. The result of the performance evaluation for *person name* entity using MUC-3 text corpus is tabled.

Table 1. Performance Evaluation Results

System	Precision	Recall	F-Measure
Look-up based NER	0.79	0.85	0.81
Stanford NER	0.34	0.17	0.23

As can be seen, the results clearly indicates that our system outperform the Stanford parser. This experiment found that the Exact Match evaluation method used in CoNLL considers correct entities only if they are exactly match with the corresponding entities in the key tests [4]. Additionally, CoNLL dataset is a collection of newswire articles of general genres. Stanford NER was also trained on the MUC dataset as mentioned earlier: 1) MUC-6 is a collection of newswire articles on negotiation of labor disputes and corporate management succession and 2) MUC-7 is a collection of newswire articles on airplane crashes, rocket and missile launches. These are copyright dataset of the North American News Text Corpora which totally different genres and structures than the MUC-3. This shows that our look-up based NER performs better in MUC-3 dataset due to the use of the regular expression rules related to terrorism texts and the *Noise Filterer* module to filter unwanted *person name.*

However, our system achieved an average level of name recognition with 79% precision, 85% recall and an F-Measure of 81% as compared to human extraction. Technically, it is clear that having smaller test data and limiting lookups to noun phrases, as opposed to sentences, is undesirable. Our data observation found that it would be impossible to identify the name “Peruvian Landazuri” and “Pelito Landazuri” from the snippet “Peruvian and Pelito Juan Landazuri”. Long names are often used by the Latin American people. In addition, due to the limited grammar rules, the system was likely to identify an organization as a person in cases such as “Adam Ltd” and “Phoenix Steel”.

6 Conclusion

A domain-specific gazetteer often relies on domain-specific knowledge to improve system performance. It is generally agreed that domain-specific entities in technical corpora, such as the terrorism one, are much harder to recognize and the results have been less satisfactory than anticipated. This is due to the built-in complexity of terms in different domains which comprise of multiword expressions, spelling variations, acronyms, ambiguities, etc. The results indicate that our proposed approach is still immature and needs further improvement. The major drawbacks, in particular, are the limited test data, the non-specific grammar rules and the multiple occurrences of names in documents. In future, improvements will be made to address these weaknesses, to increase accuracy, as well as to identify entities that are more specifically related to counterterrorism, such as *location*, *weapon*, *tactic*, *transportation*, *type of document*, etc. These also include the cardinal entities such as *date*, *time*, *number of fatalities*, *number of injuries* and *money*.

References

1. Mikheev, A., Moens, M., Grover, C.: Name Entity Recognition without Gazetteers. In: 9th Conference of European Chapter of the Association of Computational Linguistic, pp. 1–8 (1999)
2. Nissim, M., Markert, K.: Syntactic Features and Word Similarity for Supervised Metonymy Resolution. In: 10th Conference of European Chapter of the Association of Computational Linguistic, pp. 56–63 (2003)
3. Tanenblatt, M., Coden, A., Sominsky, I.: The ConceptMapper Approach to Named Entity Recognition. In: 7th Language Resource and Evaluation Conference, pp. 546–551 (2010)

4. Nadeau, D.: Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision. PhD Thesis, University of Ottawa, Canada (2007)
5. Stevenson, M., Gaizauskas, R.: Using Corpus-derived Name Lists for Named Entity Recognition. In: North American Chapter of Association for Computational Linguistics, pp. 290–295 (2000)
6. Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: 31th Conference on Computational Natural Language Learning, pp. 147–155 (2009)
7. Minkov., E., Wang, R.C., Cohen, W.W.: Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text. In: Human Language Technology / Empirical Methods in Natural Language Processing, pp. 443–450 (2005)
8. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2006)
9. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: International Conference on Computational Linguistics, pp. 539–545 (1992)
10. Riloff, E., Jones, R.: Learning Dictionaries for Information Extraction using Multi-level Bootstrapping. In: 16th National Conference on Artificial Intelligence, pp. 474–479 (1999)
11. Etzioni, O., Cafarella, M., Downey, D., Popescu, D., Shaked, A.M., Soderland, T., Weldnad, D.S., Yates, A.: Unsupervised Named Entity Extraction from the Web: An Experimental Study. J. Artificial Intelligence 165, 91–134 (2005)
12. Sasano, R., Kurohashi, S.: Japanese Named Entity Recognition using Structural Natural Language Processing. In: 3rd International Joint Conference on Natural Language Processing, pp. 607–612 (2008)
13. Pang, W., Fan, X., Gu, Y., Yu, J.: Chinese Unknown Words Extraction Based on Word-Level Characteristics. In: 9th International Conference on Hybrid Intelligent System, pp. 361–366 (2009)
14. Krieger, H.U., Schäfer, U.: DL Meet FL: A Bidirectional Mapping between Ontologies and Linguistic Knowledge. In: 23rd International Conference on Computational Linguistics, pp. 588–596 (2010)
15. Zamin, N., Oxley, A.: Information Extraction for Counter-Terrorism: A Survey on Link Analysis. In: International Symposium on Information Technology, pp. 1211–1215 (2010)
16. Zamin, N., Oxley, A.: Unapparent Information Revelation: A Knowledge Discovery using Concept Chain Graph Approach. In: National Seminar on Computer Science and Mathematics (2010) (Internal Publication)
17. Zamin, N.: Information Extraction for Counter-Terrorism: A Survey. Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 520–526 (2009)
18. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. J. Computational Linguistics 21(4), 543–556 (1995)
19. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370 (2005)

Fuzzy Goal Programming for Multi-level Multi-objective Problem: An Additive Model

Nureize Arbaiy[1,2] and Junzo Watada[1]

[1] Graduate School of Information, Production and System, Waseda University,
2-7 Hibikino, Wakamatsu, Kitakyushu, 808-0135 Japan
nureize@uthm.edu.my, junzow@osb.att.ne.jp
[2] Faculty of Computer Science and Information Technology,
University Tun Hussein Onn Malaysia, 86400 Johor, Malaysia

Abstract. The coordination of decision authority is noteworthy especially in a complex multi-level structured organization, which faces multi-objective problems to achieve overall organization targets. However, the standard formulation of mathematical programming problems assumes that a single decision maker made the decisions. Nevertheless it should be appropriate to establish the formulations of mathematical models based on multi-level programming method embracing multiple objectives. Yet, it is realized that sometimes estimating the coefficients of objective functions in the multi-objective model are difficult when the statistical data contain random and fuzzy information. Hence, this paper proposes a fuzzy goal programming additive model, to solve a multi-level multi-objective problem in a fuzzy environment, which can attain a satisfaction solution. A numerical example of production planning problem illustrates the proposed solution approach and highlights its advantages that consider the inherent uncertainties in developing the multi-level multi-objective model.

Keywords: Multi-level, multi-objective problem, fuzzy goal programming, additive model.

1 Introduction

The real situations of making decision in an organization involve a diversity of evaluation such as evaluating alternatives and attaining several goals at the same time. In many practical decision making activities, decision making structure has been changing from a single decision maker with single criterion to multiple decision makers with multi-criteria and even to multi-level situations. A resource planning problem in an organization usually consists of several objectives and requires a compromise among several committing individuals or units. Typically, these groups of decision making are arranged in an administrative hierarchical structure to supervise the independent and perhaps conflicting objectives. In this type of multi-level organization, decision planning should concern issues of central administration and coordination of decision making among lower-level activities to achieve the overall organization target. Each decision maker is responsible for one decision making unit of the hierarchical decision-making levels and controls a decision to optimize the objectives at

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 81–95, 2011.

each level. Although the execution of decision moves sequentially from an upper level to a lower level, the reaction, behavior and decision of a lower-level decision maker should affect the optimization of the decision at an upper level decision maker (Baky, 2010; Ahlatcioglu and Tiryaki, 2007; Anandalingam, 1988; Mesarovic *et al.*, 1970). Because of conflicting objectives over different levels, the dissatisfaction with the decision results is often observed among the decision makers. In such cases, a proper distribution of decision authority must be established among the decision levels for most multi-level decision situations.

A mathematical multi-level multi-objective programming has often served as a basis for structuring the underlying goals and hierarchical decision making situation of such organizations (Sinha and Sinha, 2004, Shih *et al.,* 1996). Subsequently, a multi-objective linear programming problem aims to optimize various conflicting linear objective functions simultaneously under given linear constraints to find compromise solutions (Yu, 1985 and Zeleny, 1982). Let $\mathbf{c}_i = (c_{i1}, \dots, c_{ik}), i = 1, \dots, p$ denote a vector of coefficients of the i^{th} objective function $f_i(\mathbf{x})$. Then, the multi-objective linear programming problem is written as:

$$\begin{aligned} &opt && (f_1(\mathbf{x}), \dots, f_p(\mathbf{x})) \\ &\text{s.t.} && \mathbf{Ax} \le \mathbf{b}, \\ &&& \mathbf{x} \ge \mathbf{0}, \end{aligned} \tag{1}$$

where opt indicates optimization operation (minimization or maximization), $\mathbf{x}$ is an $n-$vector with components $x_1, \dots, x_n$, $\mathbf{Ax} \le \mathbf{b}$ denotes system constraints written in vector notation and $f_i(\mathbf{x}) = \mathbf{c}_i\mathbf{x}$ are the objectives function. Nevertheless, the standard mathematical programming of multi-objective problem (1) cannot accommodate problems in a multi-level decision making structure as it is assumed that each of all objectives comes from a single decision maker at a single level. Therefore, a multi-level multi-objective programming problem solution is a necessity.

In a multi-level decision-making context, each decision maker represents a decision-making unit at a different level. All decision makers should cooperate with others in making the decision. For necessity in the sequential multi-level decision making structure, a decision maker at the highest level determines the plan and distributes this information to all decision makers in the subordinate levels. To ensure all decisions are made in cooperatively and decision authorities are distributed properly in the organization, the satisfaction of decision makers at the lower level must be considered. From this standpoint, it is desirable to develop a fuzzy programming method that facilitates multiple objectives in multi-level and fuzzy decision-making situations. In this paper, we introduce an additive model of a Fuzzy Goal Programming (FGP) approach (Tiwari *et al.,* 1987) to realize the multi-level multi-objective decision making. The FGP approach is used to achieve the highest degree of achievement for each goal by maximizing fuzzy achievement functions. The algorithm uses the concept of satisfaction to multi-objective optimization at every level until a preferred solution is attained. The problem model was also developed by means of fuzzy random regression (Nureize and Watada, 2010a) approach, to overcome the difficulties in determining the model coefficients and in treating the hybrid uncertainties that exist in the data used to construct the model coefficients. From that we emphasize that the proposed method has significant advantages in solving multi-objective problem in the multi-level organizational situation in which fuzzy random information coexisting.

The remainder of this paper is divided into six sections. Section 2 provides preliminary knowledge for a multi-level multi-objective problem and fuzzy random regression model. Section 3 explains the main components of an additive model of FGP. Section 4 describes the FGP solution algorithm for solving multi-level multi-objective problems. An illustrative example is presented in Section 5, and finally, discussions and conclusions are given in Section 6.

2 Preliminary Studies

This section explains the multi-level multi-objective decision making problem that hierarchical human organizations face to derive a rational and satisfactory solution. The decision making problems are formulated as relevant mathematical programming problems which optimization techniques can solve. Though, developing a mathematical programming model requires an appropriate model setting to avoid solutions from being mislead. Thus, the fuzzy random regression approach has been introduced in the construction of a multi-level multi-objective model.

2.1 Multi-level Multi-objective Decision Making Problem

In any organization with a hierarchical decision structure, the sequential and preemptive nature of the decision process increases complexities in making organization decision. In the multi-level programming, sequential decision making process used to start at the highest level. A decision maker at one level controls or coordinates the decision makers on the subordinate levels. Moreover, it is assumed that a decision maker at each level has a certain degree of autonomy, where the decision maker has an authority to decide the best option among the alternatives in their decision making unit. Planning in such an environment has been recognized as an important decision making process as described in Bialas and Karwan (1979).

A multi-level multi-objective programming problem is characterized when a multiple decision makers optimize several objectives in the multi-level structured organization (Sinha and Sinha, 2004; Shih *et al.*, 1996). In a multi-level programming, chosen decision variables x_{ik}^{*} are controlled by the decision maker for each level and are distributed down to the following level so that the decision-making process at the present level can include the decision from the upper level simultaneously. As each decision making level deals with several conflicting objectives, the situation creates multi-level programming problems in a set of nested optimizations over a single feasible region. In such a situation, the coordination of decision authority demonstrates that the decision variables of one level affect the decisions of the other levels. Hence, it explains that the important feature of the multi-level programming problem is essentially related to the coordination of the decision powers among all levels and that decisions at the lower levels are influenced from the upper levels.

There are many planning and/or decision making situations that can be properly represented by a multi-level programming model. All of them appear whenever a hierarchical structure is existing in the decision making process. Let us consider an organization that has a multi-level programming problem with multi-objective function $F_i(\mathbf{x})$ for $i = 1, \cdots, p,$ defined over a jointly dependent strategy set S . Let the vector of decision variables $\mathbf{x} = (x_1, \ldots, x_p)$ takes values in R^n . Assume that decisions are made sequentially beginning with DM_1 , which controls a vector $\mathbf{x}_1 \in X_1$, down through DM_p , which controls a vector $\mathbf{x}_p \in X_p$, where X_i is a nonempty subset of $\Re^{n_i}, i = 1, \cdots, p$ and $n_i = n_1 + \cdots + n_p$.

The decision maker DM_i at the i^{th} level has authority over the decision variable $\mathbf{x}_i$. The multi-level multi-objective linear programming problem is a nested optimization problem (Mohamed, 1997; Sakawa, 1993; Baky, 2010), and has the following structures:

$$
\begin{aligned}
&\text{Find } \mathbf{x} \text{ so as to} \\
&\min_{x_1 \in X_1} F_1(\mathbf{x}) = \min_{x_1} \left\{ f_{11}(\mathbf{x}), \cdots, f_{1m_1}(\mathbf{x}) \right\}, \\
&\quad \text{where } x_1 \text{ solves} \\
&\quad \min_{x_2 \in X_2} F_2(\mathbf{x}) = \min_{x_2} \left\{ f_{21}(\mathbf{x}), \cdots, f_{2m_2}(\mathbf{x}) \right\}, \\
&\qquad \vdots \\
&\qquad \text{where } x_2, \ldots, x_p \text{ solves} \\
&\qquad \min_{x_p \in X_p} F_p(\mathbf{x}) = \min_{x_p} \left\{ f_{p1}(\mathbf{x}), \cdots, f_{pm_p}(\mathbf{x}) \right\} \\
&\text{s.t.: } \mathbf{x} \in X, \\
&\quad S = \left\{ \mathbf{x} \in \Re^n : \mathbf{Ax}(\leq, \geq, =)\mathbf{b} \right\}, \\
&\quad \mathbf{x} \geq 0,
\end{aligned}
\tag{2}
$$

where $f_{ij}(\mathbf{x}) = c_1^{ij}\mathbf{x}_1 + \cdots + c_p^{ij}\mathbf{x}_p, \quad i = 1, \cdots, p, \quad j=1,\cdots,m_i,$ are linear objective functions. Let us indicate c_k^{ij} as constants, $\mathbf{A}_i$ as coefficient matrices of size $m \times n_i$ and n_i as the number of involved decision makers.

The execution of decision-making units moves from higher to lower levels. Each decision-making unit optimizes its objective function independent of other units but is affected by the actions of other level. The lower-level decision maker independently optimizes the unit's plan of action according to the goals and limitations determined in the unit, disregarding the goals of the higher-level decision maker. Thus, the problem with decision authority coordination in this multi-level structure is to identify the best compromising solution at each decision-making level to attain overall organization targets.

2.2 Multi-objective Model Setting through a Fuzzy Random Regression Approach

Typical multi-objective problem is a decision problem to optimize a set of objectives. Mathematical model is then used to represent and solve the problem. Though, the model coefficients play a pivotal role in the mathematical modeling and the value of model coefficient should be determined in prior to construct the mathematical model. The coefficients of the mathematical model are commonly decided by a decision maker with their knowledge and expertise. Nonetheless, sometimes it is not easy to determine the coefficients, as relevant data are occasionally not given or difficult to obtain. This task may cause difficulties, and thus it makes the decisions of model coefficient is crucial and influential to the model's result. The occurrence of errors in the determination of the coefficients might ruin the model formulation (Schniederjans, 1995). Therefore, a number of studies have suggested various methods to minimize these potential errors and to address the problem (Saaty, 1980; Sugihara *et al.,* 2004; Romero, 1991; Ijiri, 1968; Kwak *et al.*, 1991; Nureize and Watada, 2010b). The regression analysis will possibly work to estimate the coefficients of the model (Nureize and Watada, 2010b; 2010c).

A regression method analyzes statistical data to estimate the model coefficients in developing effective models. The conventional mathematical programming problem uses numerical deterministic values to these coefficients. In contrary, it is more realistic to take the estimated values of the coefficients as imprecise values rather than precise ones. In practical systems, probabilistic or/and vague situations include uncertain information such as predictions of future profits and incomplete historical data. Therefore, the mathematical programming models should be able to handle the above problems. That is, the above situations should be explicitly considered in the decision making process. For that reason, the fuzzy random regression model is introduced to solve such a problem with the existence of the randomness and fuzziness in historical data used for the approximation (Watada *et al.,* 2009). The property of fuzzy random regression model is used to allow for the co-existence of fuzziness and randomness in the data.

In this paper, one sigma confidence interval is used to express the confidence interval that expresses the expectation and variance of a fuzzy random variable as follows:

$$I\left[e_X, \sigma_X\right] \underline{\underline{\Delta}} \left[E(X) - \sqrt{var(X)}, \quad E(X) + \sqrt{var(X)}\right] \tag{3}$$

The fuzzy random regression model with one sigma confidence intervals (Watada *et al.,* 2009) is described as follows:

$$\begin{aligned}
&\min_{A} \; J(c) = \sum_{j=1}^{m}\left(c_j^r - c_j^l\right) \\
&c_j^r \geq c_j^l, \\
&Y_i = c_i \; I[e_{X_{i1}}, \sigma_{X_{i1}}] + \cdots + c_K \; I[e_{X_{im}}, \sigma_{X_{im}}] \underset{\tilde{h}}{\supseteq} I[e_{Y_i}, \sigma_{Y_i}] \\
&\text{for} \quad i = 1, \cdots, p \quad j = 1, \cdots, m.
\end{aligned} \tag{4}$$

where $\underset{h}{\supset}$ denotes the fuzzy inclusion at level h.

Thus, the fuzzy random regression model with confidence intervals is given in the following expression:

$$Y_i = \sum_{j=1}^{m} c_j I\left[e_{X_{ij}} + \sigma_{X_{ij}}\right], \quad i = 1, \ldots, p. \tag{5}$$

3 The Fuzzy Goal Programming Approach to the Multi-level Decision-Making Problem

Let us consider the multi-objective problem (1). In the fuzzy multi-objective problem, the objective functions are denoted as $F_i(x) \tilde{\geq} g_i$ where $\tilde{\geq}$ represents fuzzy inequality and g_i is the goal target for the objective function. Let V represent the fuzzy achievement function consisting of membership functions μ_i for fuzzy objectives. In the FGP approach, the weighted additive model (Tiwari *et al.*, 1987) is formulated by aggregating the membership functions with an additive operator as follows:

$$\begin{aligned}
\max \quad & V(\mu) = \sum_{i=1}^{m} \omega_i \mu_i \\
\text{subject to} \quad & \mu_i = \frac{\mathbf{A}_i \mathbf{X}_i - L_i}{g_i - L_i}, \\
& \mathbf{Ax} \leq \mathbf{b}, \\
& x \leq 0, \\
& \mu_i \in [0,1]; i = 1, \ldots, p.
\end{aligned} \tag{6}$$

In this section, we explain the important components required to build the additive model of FGP consisting of objective function, achievement function, goal and tolerance, and membership function.

3.1 Objective Function

The term ‘objective’ is the terminology used in goal programming approach and referred as a criterion with additional information of the direction (maximize or minimize) in which the decision maker prefers on the criterion scale (Jones and Tamiz, 2010). In a multi-objective problem, objective function $F_i(\mathbf{x})$ is created for each objective to solve. The objective function is represented in the form of

$$F_i(\mathbf{x}) = c_1\mathbf{x}_{i1} + \cdots + c_m\mathbf{x}_{im}, i = 1,\ldots,p, j = 1,\ldots,m. \tag{7}$$

In this proposed model, the coefficient value of c_{ij} is decided by the fuzzy random regression approach. . The coefficient value derived from fuzzy random regression model (4) however results in an interval denoted by the bracketed numbers $\left[c_j^l, c_j^r\right]$. Considering the midpoint value of $\xi_{ij} = \frac{\left(c^l + c^r\right)}{2}$, then the fuzzy random based objective functions (7) for FGP are rewritten as follows:

$$F_i(\mathbf{x}) = \xi_1\mathbf{x}_{i1} + \cdots + \xi_m\mathbf{x}_{im}, \quad i = 1,\ldots,p, j = 1,\ldots,m \tag{8}$$

where ξ_{ij} is the fuzzy random based coefficient.

Hence, the coefficients ξ_{ij} of each objective function are identified by the regression model and these objective functions are further used in the setting of multi-objective model.

3.2 Achievement Function

The fuzzy achievement function V is the total achievement of all the objectives. All the membership functions of the fuzzy objectives are multiplied by a weight ω that reflects their relative importance and are added together to form the achievement function.

The first level achievement function is expressed as follows:

$$\max\ V(\mu_1) = \sum_{j=1}^{m_1} \omega_{1j}\mu_{1j}. \tag{9}$$

For the subsequent lower level, the achievement function $V(\mu_p)$ is written as

$$\begin{aligned} \max\ V(\mu_p) &= \sum_{j=1}^{m_1} \omega_{1j}\mu_{1j} + \cdots + \sum_{j=1}^{m_p} \omega_{pj}\mu_{pj} \\ &+ \sum_{k=1}^{n_{(p-1)}} \left[\omega_{(p-1)k}\mu(x)_{(p-1)k}^L + \omega_{(p-1)k}\mu(x)_{(p-1)k}^R\right] \\ & k = 1,\ldots,n_i \end{aligned} \tag{10}$$

where the weight of decision variables and controlled decision vector x_{ij} is elicited with Equations (14.1) and (14.2), respectively. The weighting scheme is explained in the sub-section 3.5.

3.3 Goal and Tolerance

A goal in goal programming is known as a numerical target value that decision makers desire to achieve (Jones and Tamiz, 2010). Usually, decision makers assign values to the goal and the tolerance based on their experience and knowledge. The mathematical model can also be used to determine the goal and the tolerance values by computing the individual optimal solutions to obtain the satisfaction degree (Zimmermann, 1978).

3.4 Membership Function

Based on fuzzy set theory (Zadeh, 1965), the fuzzy objectives in a multi-objective problem are characterized by their associated membership functions. That is, the membership functions are used to formulate the corresponding objective functions. The linear membership functions μ_i for the i^{th} fuzzy objective $F_i(x) \tilde{\geq} g_i$ can be formulated as follows (Zimmermann, 1978):

$$\mu_{F_i}(x) = \begin{cases} 1 & \text{if } L_{ij} \leq f_{ij}(x) \\ \dfrac{f_{ij}(x) - L_{ij}}{g_{ij} - L_{ij}} & \text{if g}_{ij} \leq f_{ij}(x) \leq L_{ij} \,. \\ 0 & \text{if } f_{ij}(x) \leq g_{ij} \end{cases} \tag{11}$$

The membership function for $F_i(x) \tilde{\leq} g_i$ is as following:

$$\mu_{F_i}(x) = \begin{cases} 1 & \text{if } f_{ij}(x) \leq g_{ij} \\ \dfrac{L_{ij} - f_{ij}(x)}{L_{ij} - g_{ij}} & \text{if } g_{ij} \leq f_{ij}(x) \leq L_{ij} \\ 0 & \text{if } L_{ij} \leq f_{ij}(x) \end{cases} \tag{12}$$

where $i = 1,\ldots,p,\quad j = 1,\ldots,m_i$ and L_{ij} is the tolerance limit for fuzzy objectives. The membership function of each fuzzy objective was built to find the optimal solutions of the i^{th} level of the multi objective linear programming problem $x^{i*} = \left(x_1^{i*},\ldots,x_p^{i*}\right),\ i = 1,\ldots,p-1$.

In a multi-level decision-making situation, the decision at each subordinate level influences the upper level's decision as well as the upper-level decision makers control the subordinate level's decision. The decision denoted as x_{ik} in the present level is sent down to the next lower level. To take care of the vagueness of this decision x_{ik} , let $t_k^{i_L}$ and $t_k^{i_R}$ for $i = 1,\ldots, p-1;\quad k = 1,\ldots,n_i$ be the maximum negative and positive of tolerance values, respectively, for the decision vectors x_{ik} with values specified by the i^{th} level decision maker.

The triangular fuzzy numbers of the decision vectors x_{ik} are stated as $\left(x_{ik}^* - t_k^{i_L}, x_{ik}^*, x_{ik}^* + t_k^{i_R}\right)$. Thus, as in Baky (2010), the linear membership functions for each of the n_i components of the decision vector $x_i^{i*} = \left(x_1^{i*},\ldots,x_p^{i*}\right)$ controlled by the decision makers of the upper $p-1$ levels can be formulated as follows:

$$\mu_{x_{ik}}(\mathbf{x}) = \begin{cases} \dfrac{x_{ik} - \left(x_{ik}^* - t_k^{i_L}\right)}{t_k^{i_L}} & \text{if } x_{ik}^* - t_k^{i_L} \le x_{ik} \le x_{ik}^* \\ \dfrac{\left(x_{ik}^* + t_k^{i_R}\right) - x_{ik}}{t_k^{i_R}} & \text{if } x_{ik}^* \le x_{ik} \le x_{ik}^* + t_k^{i_R} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $i = 1,\ldots, p-1;\quad k = 1,\ldots,n_i$.

3.5 Relative Importance

Let the numerical coefficients ω_{ij}^+ , ω_{ik}^R and ω_{ik}^L denote the relative importance of achieving the aspired levels. The relative importance of the fuzzy goal is then determined using the weighting scheme (Mohamed, 1997).

$$\omega_{ij}^+ = \frac{1}{u_{ij} - g_{ij}}, \tag{14.1}$$

$$\omega_{ik}^L = \frac{1}{t_k^{i_L}},\quad \omega_{ik}^R = \frac{1}{t_k^{i_R}}. \tag{14.2}$$

ω_{ij}^+ is the weight for the objective functions, and ω_{ik}^R and ω_{ik}^L represent the weights for the membership functions of the decision vectors.

4 Fuzzy Goal Programming for the Multi-level Decision Problem: An Additive Model

A solution to the multi-level multi-objective programming problem is obtained as follows on the basis of the main components of additive FGP. For two continuous levels in the decision making tree, the decision-making process is carried out in two sequential stages. The higher level decision maker determines the top plan of action and followed by the lower level decision maker that executes the plan which is decided by the higher level decision maker.

The additive FGP model for multi-level decision making is written as follows:

$$\begin{aligned}
\max\ V(\mu_p) &= \sum_{j=1}^{m_1}\omega_{1j}\mu_{1j}+\cdots+\sum_{j=1}^{m_p}\omega_{pj}\mu_{pj}\\
&\quad+\sum_{k=1}^{n_{(p-1)}}\left[\omega_{(p-1)k}\mu(x)^L_{(p-1)k}+\omega_{(p-1)k}\mu(x)^R_{(p-1)k}\right]\\
\text{subject to}:\quad \mu_{\mathrm{F_i}(\mathbf{x})} &= \frac{\left(c_1^{ij}x_1+\cdots+c_p^{ij}x_p\right)-\mu_{ij}}{g_{ij}-\mu_{ij}},\quad i=1,\ldots,p,\ j=1,\ldots,m_i,\\
\mu^L_{x_{ik}}(\mathbf{x}) &= \frac{x_{ik}-\left(x^*_{ik}-t_k^{i_L}\right)}{t_k^{i_L}},\quad i=1,\ldots,p-1,k=1,\ldots,n_i,\\
\mu^R_{x_{ik}}(\mathbf{x}) &= \frac{\left(x^*_{ik}-t_k^{i_R}\right)-x_{ik}}{t_k^{i_R}},\quad i=1,\ldots,p-1,k=1,\ldots,n_i,\\
&A_i x_i(\le,\ge,=)b,\quad \mathbf{x}\ge 0,\\
&\mu_i\in[0,1],\quad i=1,\ldots,m.
\end{aligned}\tag{11}$$

Fig 1 illustrates the process flow of decision making under multi-level additive fuzzy goal programming model. The multi-level additive FGP is separately solved for the i^{th} level multi-objective program with $i=1,\ldots,p-1$.

5 An Illustrative Example

Consider the following production planning problem. One export-oriented country is concentrating on producing three important products x_1 , x_2 and x_3 which are manufactured by company C_d, $d=1,\cdots,D$ with given capabilities. This company has distributed branch B_d, $d=1,\cdots,D$, in city level for producing the products. This situations result in 3 levels decision making and each level is responsible to accomplish the objectives that are decided in prior.

The initial step in this phase is the data preparation to determine the decision variable's coefficient through a Fuzzy Random Regression Model (4) and further develop the objective function for multi-level multi-objective problem (14). The previously collected data set is then pre-processed (Watada *et al.,* 2009; Nureize and Watada, 2010c). The probabilities are assigned as the proportion of product being produced in i^{th} plant to the total production numbers.

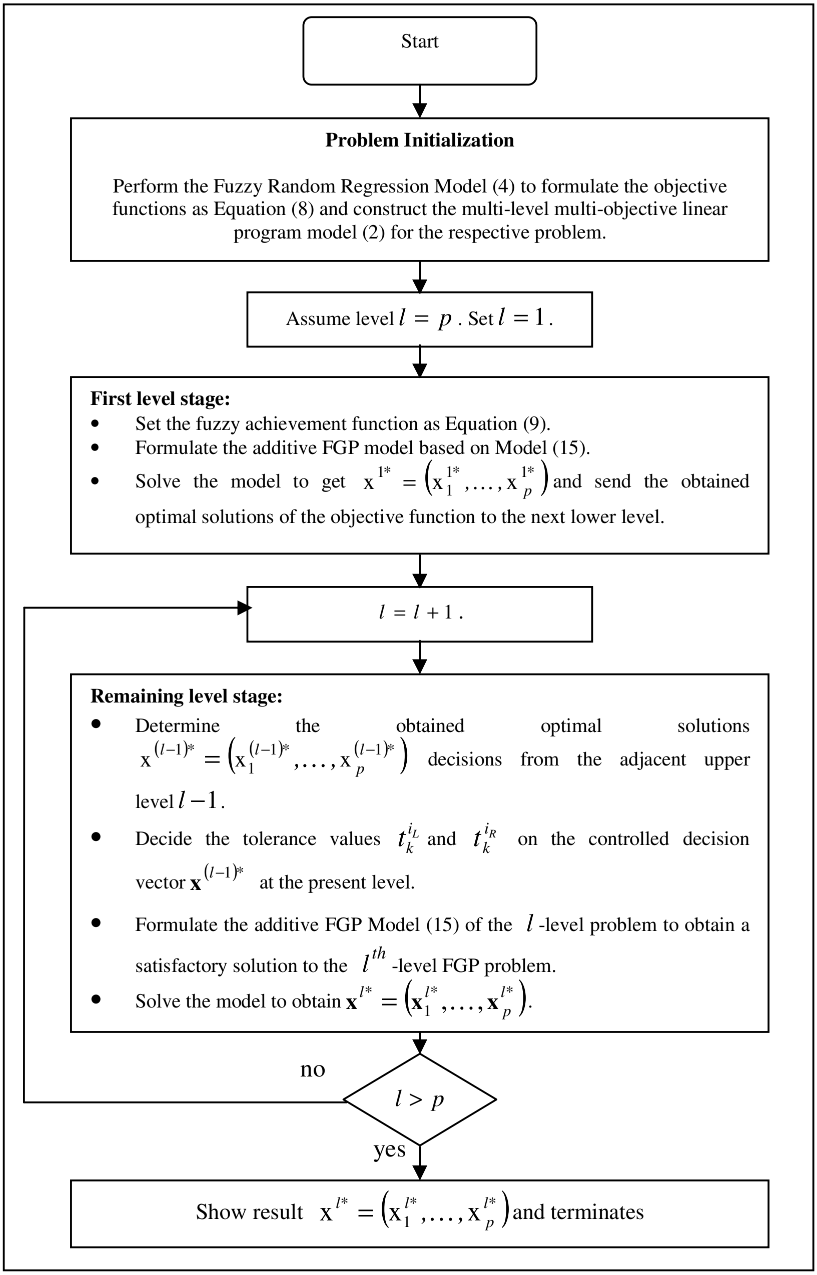

Start
Problem Initialization
Perform the Fuzzy Random Regression Model (4) to formulate the objective functions as Equation (8) and construct the multi-level multi-objective linear program model (2) for the respective problem.
Assume level $l = p$. Set $l = 1$.
First level stage:
• Set the fuzzy achievement function as Equation (9).
• Formulate the additive FGP model based on Model (15).
• Solve the model to get $x^{1*} = \left(x_1^{1*}, \dots, x_p^{1*}\right)$ and send the obtained optimal solutions of the objective function to the next lower level.
$l = l + 1$.
Remaining level stage:
• Determine the obtained optimal solutions $x^{(l-1)*} = \left(x_1^{(l-1)*}, \dots, x_p^{(l-1)*}\right)$ decisions from the adjacent upper level $l - 1$.
• Decide the tolerance values $t_k^{i_L}$ and $t_k^{i_R}$ on the controlled decision vector $\mathbf{x}^{(l-1)*}$ at the present level.
• Formulate the additive FGP Model (15) of the l -level problem to obtain a satisfactory solution to the l^{th} -level FGP problem.
• Solve the model to obtain $\mathbf{x}^{l*} = \left(\mathbf{x}_1^{l*}, \dots, \mathbf{x}_p^{l*}\right)$.
no
$l > p$
yes
Show result $x^{l*} = \left(x_1^{l*}, \dots, x_p^{l*}\right)$ and terminates

Fig. 1. Flowchart of the FGP for Multi-level Multi-Objective Problem

Table 1 summarizes the information needed to construct the multi-level multi-objective model (2). Let us assume f_{ij} represents the DM_i objective(s) in each level. Based on the information in Table 1, the multi-level multi-objective problem can be summarized as follows:

Find x so as to satisfy:

$$\begin{array}{l}
\text{[Level1]}\ \min\limits_{x_1} \\
\begin{pmatrix} 2.078x_1 + 0.260x_2 + 0.170x_3 \geq g_{11}, \\ 1.010x_1 + 1.700x_2 + 0.476x_3 \geq g_{12}, \\ 0.438x_1 + 0.796\,x_2 + 0.512x_3 \leq g_{13}. \end{pmatrix} \\
\text{[Level2]} \\
\text{where}\, x_2 \text{ and}\, x_3 \text{ solves,} \\
\quad \min\limits_{x_2} = \\
\quad \begin{pmatrix} 1.126x_1 + 0.100x_2 + 0.097x_3 \geq g_{21}, \\ 0.856x_1 + 1.473x_2 + 0.443x_3 \geq g_{22}, \\ 0.380x_1 + 0.737x_2 + 0.277x_3 \leq g_{23}. \end{pmatrix} \\
\text{[Level3]} \\
\quad \text{where}\, x_3 \text{ solves,} \\
\quad\quad \min\limits_{x_3} \\
\quad\quad \begin{pmatrix} 0.921x_1 + 0.050x_2 + 0.526x_3 \geq g_{31}, \\ 0.380x_1 + 0.737x_2 + 0.216x_3 \leq g_{32}. \end{pmatrix}
\end{array} \tag{14}$$

Subject to constraints:

$$\begin{array}{ll}
\text{raw} & 3.815x_1 + 0.910x_2 + 0.220x_3 <= 87.75; \\
\text{labor} & 0.650x_1 + 0.900x_2 + 0.125x_3 <= 4.425; \\
\text{mills} & 17.50x_1 + 2.160x_2 + 4.775x_3 <= 95.20; \\
\text{capital} & 1.350x_1 + 0.980x_2 + 0.890x_3 <= 20.15;
\end{array} \tag{15}$$

Note that all the coefficients for the objective functions in the problem model (14) are derived from Fuzzy Random Regression Model (4).

Based on the procedure stated in Section 4 and the workflow in Fig. 1, three equivalent linear programs are constructed in sequence. Table 2 tabulates the goal and tolerance that are pre-determined by the experts for all objectives functions of the three levels of the multi-level multi-objective problem. Computer software LINGO© is used to solve the equivalent ordinary linear programming model. The procedure brings to an end as $l = p = 3$ and the satisfactory solution is obtained.

6 Discussions and Conclusions

This section is spent to clarify the results of the production planning problem. The Fuzzy Random Regression Model (4) gave the coefficients values in interval value form, which expresses fuzzy judgments. In this case, first stage of computation used the midpoint values of the interval coefficients. The coefficients ξ_i are fuzzy random coefficients for the decision variables x_i as tabulated in Table 1. The optimal solution for each level with the controlled decision variables for the problem is obtained as in Table 2.

Table 1. The coefficients for objective functions and goal's target

Decision Making Level		Goal	Fuzzy Random-Based Coefficient			Target	Tolerance
			ξ_1	ξ_2	ξ_3		
Government level, DM_1 (first-level)	f_{11}	Maximize the export revenue	2.078	0.260	0.170	4.58	0.5
	f_{12}	Maximize the national level profit	1.010	1.700	0.476	5.50	0.5
	f_{13}	Minimize the capital,	0.438	0.796	0.512	5.00	0.5
State level, DM_2 (second-level)	f_{21}	Maximize the production volume	1.126	0.100	0.097	3.90	0.5
	f_{22}	Maximize the profit for state level	0.856	1.473	0.443	4.50	0.5
	f_{23}	Minimize the cost of production	0.380	0.737	0.277	4.80	0.5
City level, DM_3 (third-level)	f_{31}	Maximize the production volume	0.921	0.050	0.526	3.00	0.5
	f_{32}	Minimize the cost of production	0.380	0.737	0.216	4.00	0.5

Table 2. The optimal solutions and decision tolerance

Decision Making Level	Solutions $x = \{x_1, x_2, x_3\}$	Controlled decision variables	Controlled decision variables tolerance	
			$t_k^{i_L}$	$t_k^{i_R}$
1	$x^{1*} = \{1.94, 2.08, 0.00\}$	x_1^{1*}	0.5	0.5
2	$x^{2*} = \{1.94, 1.92, 0.00\}$	x_2^{2*}	0.75	0.25
3	$x^{3*} = \{1.94, 1.92, 0.02\}$	-	-	-

The experiment's results show that the administrative government level (first level), objectives f_{11} and f_{21} attained nearly full satisfaction achievements which were 98% and 94%, respectively. However, the objective of minimizing the capital only partly achieved about 42% in this level. The objective to maximize the profit at the government state level has fully achieved, whereas the other objectives gained 55% and 38%. In the city level, the objectives satisfied about 55% and 47% achievements. The results show that decision makers can perform decision analysis under consideration of the solution derived by the mathematical approach. The decision maker can re-examine the solution and change the decision and repeat the process. Since the proposed method is based on the satisfaction approach, the decision makers may involve themselves in evaluating the results to attain better satisfying solution.

In this study, we demonstrated the use of the additive model of an FGP approach to solve multi-level multi-objective programming problems, where the initial problem model was developed in terms of a fuzzy random regression approach to treat the uncertainties in the data and to overcome the difficulties in determining the coefficients values. In summary, the proposed procedures have properly used the additive method in the FGP evaluation to solve multi-level multi-objective problems. The procedure also enables the decision maker of a respective decision-making unit to decide the decision value by means of the mathematical based on their satisfaction. Although it is an iterative process, it is practical for the decision maker to re-evaluate the results to attain the satisfaction of the overall system target. In addition, the decision maker's preferences toward the goals are considered in the computation of the decision process by introducing the relative importance evaluation in the additive FGP model.

Acknowledgments. A. Nureize would like to express her appreciation to the University Tun Hussein Onn Malaysia (UTHM) and the Ministry of Higher Education Malaysia (MOHE) for a study leave.

References

1. Abo-Sinna, M.A.: A Bi-Level Non Linear Multiobjective Decision-Making Under Fuzziness. Journal of Operational Research Society of India (OPSEARCH) 38(5), 484–495 (2001)
2. Ahlatcioglu, M., Tiryaki, F.: Interactive Fuzzy Programming For Decentralized Two-Level Linear Fractional Programming (DTLLFP) problems. Omega 35(4), 432–450 (2007)
3. Anandalingam, G.A.: Mathematical Programming Model of Decentralized Multilevel Systems. Journal of Operational Research Society 39(11), 1021–1033 (1988)
4. Baky, I.A.: Solving Multi-Level Multi-Objective Linear Programming Problems Through Fuzzy Goal Programming Approach. Applied Mathematical Modeling 34(9), 2377–2387 (2010)
5. Bialas, W.R., Karwan, M.H.: Mathematical Methods For Multilevel Planning, Research Report No. 79-2 (1979)
6. Ijiri, Y.: An Application of Input-Output Analysis To Some Problems In Cost Accounting. Management Accounting 15(1), 49–61 (1968)

7. Jones, D., Tamiz, M.: Practical Goal Programming. International Series in Operations Research and Management Science, vol. 141. Springer, Heidelberg (2010)
8. Kwak, N.K., Schniederjans, M.J., Warkentin, K.S.: An Application Of Linear Goal Programming To The Marketing Distribution Decision. European Journal of Operational Research 52(3), 334–344 (1991)
9. Mesarovic, M., Macko, D., Takahara, Y.: Theory of Hierarchical Multi-Level Systems. Academic Press, New York
10. Mohamed, R.H.: The Relationship between Goal Programming and Fuzzy Programming. Fuzzy Sets and Systems 89(2), 215–222 (1997)
11. Arbaiy, N., Watada, J.: Constructing Fuzzy Random Goal Constraints for Stochastic Fuzzy Goal Programming. In: Huynh, V.-N., Nakamori, Y., Lawry, J., Inuiguchi, M., et al. (eds.) Integrated Uncertainty Management and Applications. AISC, vol. 68, pp. 293–304. Springer, Heidelberg (2010a)
12. Nureize, A., Watada, J.: A Fuzzy Regression Approach To Hierarchical Evaluation Model For Oil Palm Grading. Fuzzy Optimization Decision Making 9(1), 105–122 (2010b)
13. Nureize, A., Watada, J.: Approximation of goal constraint coefficients in fuzzy goal programming. In: 2010 Second International Conference on Computer Engineering and Applications, vol. 1, pp. 161–165 (2010c)
14. Romero, C.: Handbook of critical issues in goal programming, pp. 67–71. Pergamon Press, Oxford (1991)
15. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
16. Sakawa, M., Nishizaki, I.: Interactive Fuzzy Programming For Two-Level Linear Fractional Programming Problem. Fuzzy Sets and Systems 119(1), 31–40 (2001)
17. Sakawa, M.: Fuzzy Sets and Interactive Multi-Objective Optimization, Applied Information Technology. Plenum, New York (1993)
18. Schniederjans, M.J.: Goal Programming, Methodology and Applications. Kluwer, Boston (1995)
19. Shih, H.-S., Lai, Y.-J., Lee, E.S.: Fuzzy approach for multi-level programming problems. Computers Operations Research 23(1), 73–91 (1996)
20. Sinha, S.B., Sinha, S.: A Linear Programming Approach for Linear Multi-Level Programming Problems. Journal of the Operational Research Society 55(3), 312–316 (2004)
21. Sugihara, K., Ishii, H., Tanaka, H.: On Conjoint Analysis by Rough Approximations Based On Dominance Relations. International Journal of Intelligent System 19(7), 671–679 (2004)
22. Tiwari, R.N., Dharmar, S., Rao, J.R.: Fuzzy Goal Programming An Additive Model. Fuzzy Sets and Systems 24(1), 27–34 (1987)
23. Wang, H.-F., Huang, Z.-H.: Top-Down Fuzzy Decision Making With Partial Preference Information. Fuzzy Optimization and Decision Making 1(2), 161–176 (2002)
24. Watada, J., Wang, S., Pedrycz, W.: Building Confidence-Interval-Based Fuzzy Random Regression Model. IEEE Transactions on Fuzzy Systems 11(6), 1273–1283 (2009)
25. Yu, P.L.: Multiple Criteria Decision Making: Concepts, Techniques and Extensions. Plenum, New York (1985)
26. Zadeh, L.A.: Fuzzy Sets. Information Control 8(3), 338–353 (1965)
27. Zeleny, M.: Multiple Criteria Decision Making. McGraw-Hill, New York (1982)
28. Zimmermann, H.-J.: Fuzzy Programming and Linear Programming with Several Objective Functions. Fuzzy Sets and Systems 1(1), 45–55 (1978)

Software Agents for E-Commerce Data Workflow Management

Faiz Al-Shrouf[1], Aiman Turani[2], and Khalil Al-Shqeerat[3]

[1] Applied Science University, Faculty of Information Technology,
Department of Computer Science, 11931 Amman, Jordan
faiz_alshrouf@asu.edu.jo

[2] Applied Science University, Faculty of Information Technology,
Department of Software Engineering, 11931, Amman, Jordan
aimant@asu.edu.jo

[3] Applied Science University, Faculty of Information Technology,
Department of Computer Network Systems, 11931, Amman, Jordan
dr_khalil@asu.edu.jo

Abstract. Software agent technology started to have a key role in e-commerce domain. Agents are now used to support a pervasive technology for different partners in e-business environment in which various virtual business processes will be incorporated and facilitated. Agent technology is used to automate these processes, as well as to enhance e-market places where sellers, vendors, and retailers provide a virtual shop for consumers and buyers to purchase items online through delegating requirements to software agents. This paper proposes a new framework for the use of software agent technology. The paper presents underlying framework, which is implemented by using agent coordination and collaboration in a distributed computing environment. The use of an agent controller pattern will provide robustness and scalability to e-market place. The framework allows multiple sellers to be registered, whereas buyers satisfy their requirements by using a mobile purchasing agent, which translates their requirements to the e-market place. In addition, the framework customized to satisfy e-business transactions for buyers and sellers.

Keywords: Workflow Management, Software Agents, Consumer Buying Behavior (CBB) Model.

1 Introduction

Electronic commerce entails Business-to-Business, Business-to-Customer, and Customer-to-Customer transactions. It encompasses a wide range of issues including security, trust, reputation, payment mechanisms, advertising, electronic product catalogs, intermediaries, multimedia, shopping experience, back office management, workflow management, supply chain management, service discovery, knowledge management, automated negotiation and pricing, auctioning, and transactional reasoning [1][10][17].

It is generally accepted that there are wide range of accepted technologies used in promoting virtual businesses processes, specifically communication and networking.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 96–106, 2011.

However, medias and products have been gradually moved to e-commerce. Business partners are pushed to learn technologies in order to enhance their virtual business processes. We may predict that this new method of doing e-business will dominate business settings in the near future.

One of the major technologies used are intelligent software agents. Agent-based technologies play an important role for doing e-business processes [6] [7]. Agent-based commerce is becoming reality and many business partners acknowledge several of these technologies in their web-based business as they want to be part of the next generation of Internet commerce.

2 Intelligent Software Agents

Intelligent software agents are used to support B2B, B2C, and C2C paradigms. Basically, agents are also called "Agent-Mediated Commerce" [17], which is widely used to perform tasks such as matchmaking, monitoring, negotiation, bidding, auctioning, and transfer of goods. The role of agent-based commerce is to aid comparison in shopping process. These agents collect information from multiple commercial sites[2], filtering it and provide appropriate requests for both buyers and sellers.

Agent derives from the concept of agency, referring to employing someone to act on your behalf. There are almost as many definitions for the term software agents as there are people employing it. For examples, AIMA (Artificial Intelligence Modern Approach) [18] [19] has defined an agent as anything that can be viewed by perceiving its environment through sensors and acting on that environment through effectors. The problem with this definition is (, as argued in [14] and [15],) the need to define the environment for whatever that provides input and recieves output. In addition, there is a requirement to consider the input to be sensed and the output to be acted to make a program as an agent.

Another definition [1], an agent is defined as a software entity that functions continuously and autonomously in a particular environment, which is often inhabited by other agents and processes. Agents should perform activities in flexible way as well as responsive to changes in the environment without human supervision. Besides, an agent that functions over a long time period should be able to adopt from its experience. Agents also communicate and coordinate between each of them. [5] defined autonomous agents as computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment and by doing so realize a set of goals or tasks for which they are designed.

2.1 Agent Characterization

Generally, an agent is an entity, which is responsible for satisfying specific goals (achieving a specific status, maximizing a given function [15]), and has mental properties such as state, knowledge, belief, intension, obligation [12] [15].The "state" of an agent includes the representation of its environment and the agent "knowledge" defines the information that an agent considers as sure and agent "beliefs" represents the information that may possibly wrong. The following are characterizations of agents:

1- Situated: An agent situated in an environment that consists of objects and other agents it is possible to interact with [14].

2- Autonomous: agents can operate without human intervention (goal directedness) from others (humans or other software processes [1] and that should have control over its own actions and state [8] [14].

3- Flexibility: agents should be flexible.
4- Responsive: agents should perceive their environment and respond to changes that occur in it [14] [18].
5- Reactive: the ability to observe and sense the status of objects and agents in its environment including events that happen in it, such change the state of an object, reception of a message [18].
6- Proactive: agents should be able to take initiative when appropriate [8] [14] and when attempting to achieve goals.
7- Social: agents should be able to make interactions when appropriate with humans and other agents [15]. They also engage in planning with other agents [1].

8- Learning: agents should learn as they react and interact with external environment over time and their performance should increase [1] [18].
9- Mobility: agents should have the ability to move from one machine to another across different system architectures and platforms [1] [12].
10- Veracity: agents do not knowingly communicate false information.
11- Benevolence: agents always try to do what they are asked for [8].
12- Rationality: agents will try to achieve their goals and not act in such away to prevent their goals from being achieved [12] [14].

2.2 Mobile Software Agents

Agents are either static agents or mobile agents. Static agents or stationary agents [9] execute only on a single system during its life cycle. The life cycle of software agent comprises a life cycle model, computational model, a security model, and a communication model. A mobile agent additionally defines a navigation model. Services to create, destroy, suspend, resume, and stop agents are needed to support the agent life cycle model [9].

Mobile agents are computational software processes capable of roaming Wide Area Network (WAN), interacting with foreign hosts, gathering information on behalf of its owner and coming ‘back home’ having performed the duties set by its user. Mobile agents ensure autonomy, cooperation, and mobility, they are handling routine tasks, searching for information, and facilitating decision supporting on user behalf.

3 The Consumer Buying Behavior (CBB) Model of E-Commerce

To categorize how agents are being used in B2C paradigm, it is important to explore roles of agents as mediators in e-commerce in the context of a common model. This model stems from traditional marketing Consumer Buying Behavior (CBB) model. The CBB model is a powerful tool to help understanding the roles of agents. CBB model addresses six stages that also elucidate where agent technologies apply to the shopping experience. These stages are given below:

1-Need Identification: this is the stage where customers conceptualize a need for a product or a service. Within this stage the consumer can be stimulated through product information [17]. Agents can play an important role for those purchases that are repetitive (supplies) or predictable (habits). One of the oldest and simplest examples of software agents are called "Monitors" continuously running programs which monitor a set of sensors or data streams and take action when a certain pre-specified condition apply. Another example called "Notification Agent" or "Eyes" by "Amazon which monitors the catalog of books for sale and notifies the customer when certain events occur that may be of interest to the customer or when a new category becomes available [1].

2- *Product Brokering*: the stage where the customer determines what he / she need to buy [10]. This enables the selling agents to answer with related products and or alternatives according to the preferences set by the user [13]. The buyer has identified a need to make a purchase (possibly with the assistance of monitoring agent). There are several agent systems that reduce consumers search cost when deciding which products best meet their needs [1]. The result of this stage is a set of products [17].

3- *Merchant Brokering:* this stage combines the consideration set from the previous stage with merchant–specific information to help determine who to buy from [1] [17]. This includes the evaluation of merchant alternatives based on consumer provided criteria (e.g. price, warranty, availability, delivery time, reputation, etc…) [17]. The problem that was exposed, most of the merchants do not want to compete on price only, and want the value added services mentioned previously to be included in consumers buying decision [1].

4- *Negotiation:* is the stage where the customer may interact with the service product provider to determine the terms of transaction (i.e. price, quantity, quality of service, etc…) [11]. Negotiation in real-world business, increases transaction costs that may be too high for either consumers or merchants; also, in the real world there are impediments to using negotiation such as time constraints. These mostly disappear in digital world. In traditional retail markets, prices and other transactions are fixed. The benefit of dynamically negotiation the price for a product instead of fixing it is that it relieves the merchant from needing to determine the value of the good a priory [1].

5- *Payment and Delivery:* this is the stage where a consumer can specify the terms of how to receive the product or service [10]. The purchase and delivery of a product can either signal the termination of the negotiation stage or occur sometime afterwards. In some cases, the available payment (e.g. cash only) or delivery options can influence product and merchant brokering [17].

6- *Product Service and evaluation:* this post-purchase stage involves product service, customer service an evaluation of the satisfaction of the overall buying experience and decisions [17].

Given the above stages, there was an emphasis on agents to only support the mediation- based stages (i.e. product brokering, merchant brokering and negotiation). Currently [11] all stages are being implemented with intelligent agents. The need identification stage has been incorporated into recommended systems. These systems have the ability to analyze historical purchase data and recommend solutions to customers. Product suppliers can keep preference information on their customers and notify them when relevant product / service are available. Agents have been incorporated into recommended services for both seller and buyer. These agents can analyze market trends and determine if the users are getting reliable information or best deal for their on-line transactions [10].

Current product brokering and merchant brokering systems take large amounts of product data and help the user to narrow down the selection. Traditionally, agents have been incorporated in these environments to do tasks of browsing and comparing. In addition, there is also a huge amount of work in negotiation, auctioning and reasoning [3] [11]. However, the evaluation stage still has not seen significant reported implementations with agents [10].

4 The Proposed Framework

Buyers and sellers conduct business online through the e-market place. The market place is a place where people meet for the purpose of trade by private purchase and sale. By the invention of technologies and communications, business partners conduct their processes online. The benefit of using e-markets stems from providing physical business transactions using the Internet. As described in [16], Pedro et. al. (1999) suggests a framework for virtual market place. Their framework allows buyers and sellers to corporate business processes through using static agents-mediated e-commerce. In our proposed framework, mobile purchasing agent, which cooperates with a set of static coordination agents The framework highlights stages of CBB model mentioned earlier: need identification, product brokering, merchant brokering, and negotiation. Furthermore, the framework proposes a new analysis stage of product evaluation. Figure (1) depicts the framework for B2C e-commerce data workflow management.

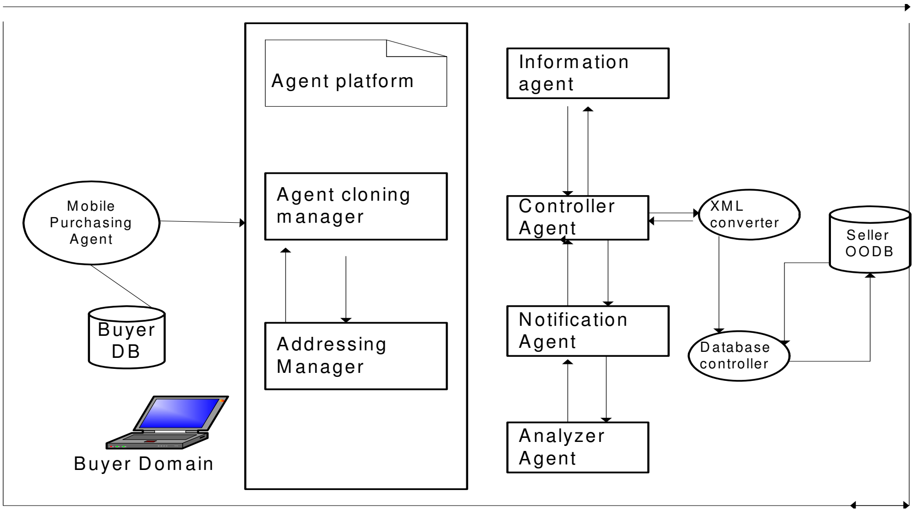


Fig. 1. Agent-Based Framework for E-Commerce Data Workflow Management

The framework proposes three main phases, the requirements phase, the mediation phase, and determination phase. These phases are described below:

4.1 Requirement Phase

The requirements phase addresses two stages of the CBB model; need identification, and product brokering. It is instantiated by allowing the buyer to identify his/her requirements in application running on a web browser. These requirements are: service name, price, quantity, and priorities. The buyer is accepting these requirements and packages them in XML format[1]. XML stands for (eXtensible Markup Language). XML is a Markup Language for describing structured data and it used to solve incompatible formats and inconsistencies between different computer systems and databases over the Internet. XML was designed for electronic document exchange. Business partners rely heavily on flow of agreements expressed in XML documents. XML is a suited platform to do business. The buyer stimulates the mobile purchasing agent, which can be created using Java2 platform for building agents. The Aglet platform from IBM is a good platform for building mobile agents. Aglet can be viewed as agent and applet or mobile applet. In our model, the mobile purchasing agent travels across the network and reaching the server site at the e-market place to perform tasks upon buyer's behalf. The mobile purchasing agent is programmed to wait till it receives results from the market.

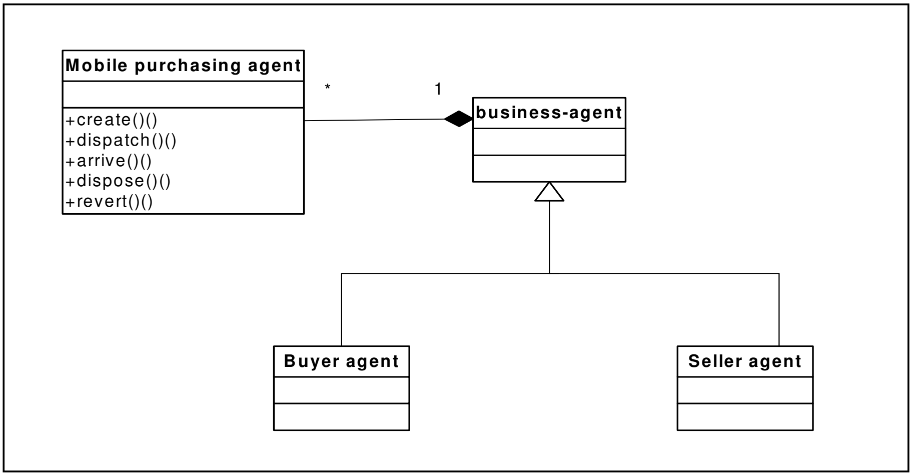


Fig. 2. UML Class Diagram of Mobile Purchasing Agent

[1] XML is defined by the World Wide Web Consortium (W3C) to be an application and vendor neutral, which insures maximum portability.

Basically, the framework intended to use a mobile purchasing agent, which translates buyer requirements. The mobile purchasing agent travels to the e-market place to co-operate with agents in agent platform. Activities of mobile agents created in IBM's Aglet or General Magic's Odyssey [9], are event driven. This means, that when a certain event occurs as detected by some "Listeners", the specific actions for that event will be executed. Figure (2), presents UML mobile purchasing agent class as a subclass of the super class business-agent.

4.2 Mediation Phase

This is the primary phase in the framework. It addresses merchant brokering and negotiation of the CBB model. The mediation phase is instantiated when the mobile purchasing agent arrives the e-marketplace. The mobile purchasing agent registers itself in agent server, which holds agents according to an agent solver manager. The components of agent solver manager are given as follows:

1- Agent cloning manager: Contents of mobile purchasing agent have to be cloned to ensure that a virus will not infect its contents when it travels across the network.
2- Addressing manager: This manager contains all addresses (URLs) from sellers in market sites. Each of which represents a seller address given through his/her URL.
3- Agent manager: The agent manager consists of a set of agents that represent the core-agents. Agents are responsible for searching and negotiation mechanisms.

- Information agent: This agent is a static agent. It is responsible to detect arrival of the mobile purchasing agent and extract the XML document and passes it to the controller agent.
- Controller agent: This agent is a static agent. It is responsible to control processes to other agents, collects the request from the information agent, passes this information to notification agent, and connects with the XML converter.
- Notification agent: This agent is responsible to conduct a notification mechanism for seller's agents. When the negotiation is finalized a deal, both the buyer and supplier are informed.
- Analyzer agent: When the negotiation is processed between supplier agent and buyer agent, the buyer and the (winner) seller are informed. The buyer and supplier send their confirmation or non-confirmation via an analyzer agent.

4- XML converter: Every market site has an XML format converter that converts an XML document between formats followed by the buyer and the local database schema and vice versa using XSL (eXtensible Style Language)[9]. The XSL provides rule for displaying and organizing XML data.
5- Database controller: The database controller receives the set of buyer requirements and constraints to form SQL query and connects with the database. The database controller receives results from the database and updates it.

4.3 Determination Phase

At this phase, the buyer has determined his choice and deal. Results matching the buyer's requests are returned from the database controller and are converted to the XML format followed at the buyer end. This resulting XML document which is transferred to the mobile purchasing agent, which waits till it receives the results from the market site and directs these results to the buyer back end.

5 Application Scenario

We implement an application in e-market place that utilizes the following agents:

1- Master Administrator Agent (Mobile Purchasing Agent): This agent requests a deal and handles message results from seller domain.
2- Model Slave Agent(Information Agent): This agent is responsible for collecting information from buyer agent domain
3- View Controller Slave Agent: (Controller Agent): This agent controls the Seller's price of a deal
4- View Agent (Notification and Analyzer Agent): This agent finalized the deal and demonstrates statistical charts representing sales.

Figure (3), Figure (4)) shows the collaboration of different agents in the application scenario and sample screenshot of view and analyzer agent respectively.

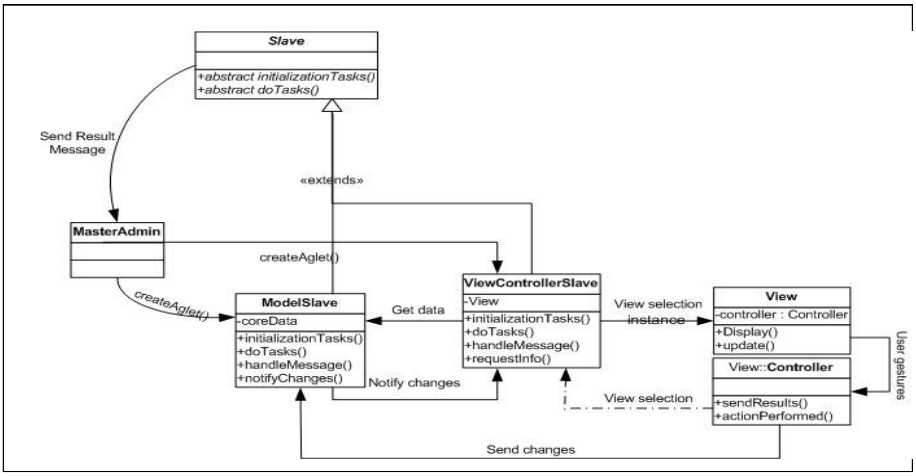


Fig. 3. Structure of Agent's Collaboration in E-Marketplace

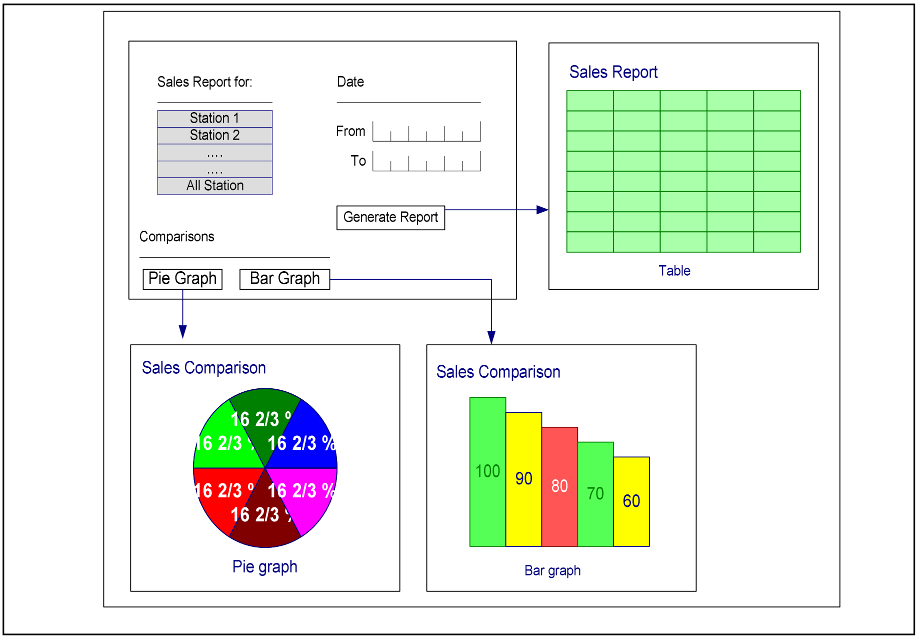


Fig. 4. Sample Screenshots of Analyzer and Notification Agents

6 Conclusions and Future Works

We have described a new framework for the Business-to-Customer paradigm. The description also entails features, components, benefits, and restrictions of the framework.

We concentrate on using intelligent software agents as a key to facilitate engineering business processes. Collaboration and coordination between mobile purchasing agent and static agents in e-market place have also been described.

The proposed framework has certain features for supporting e-market place effectiveness. These features stem from using a mobile purchasing agent, which collaborates and coordinates with static agents in the e-market site. The strength ness of the framework is due to the following:

1. Supporting an effective mechanism for coordination and collaboration between different agents in e-market sites.
2. Providing security and authentication disciplines for both buyers and sellers. Trustworthy issues are pushed to facilitating the underlying processes.
3. Using an agent controller forces our framework and provides a minor platform. This platform supports agents, which can be considered as collaborating components in business environment.
4. The agent platform maintains the contents of the mobile purchasing agent, which constitutes the basic infrastructure of the framework.

Lastly, the following restrictions and limitations might be considered:

1. Development of the framework requires standard communications and protocols between business and agent software components.
2. The framework involves standard services in e-market place, such as buying/selling workflow management processes. Other business services were not involved. Framework interoperability is not dominated CBB stage (product service and payment delivery).

Future direction for the proposed framework requires the existence of new business services and other workflow management business processes. If new services are encountered, then the framework could be improved to involve recommendations and solutions for inquiries instantiated from both buyers and sellers.

Acknowledgment

We would like to thank the Department of Computer Science, Applied Science University for supporting this research.

References

1. Pivk, A., Gams, M.: Intelligent Agents in E-Commerce. Jamova 39, 1000 Ljubljana (1999)
2. Sahuguet, A.: About agents and databases. Technical report, CIA-650 (1997)
3. Chavez, Maes, A., Kasbah, P.: An Agent Marketplace for Buying and Selling Goods. In: Proceeding of the first International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology, PAAM 1996 (1996)
4. Deitel, Nieto: E-business & E- commerce: How to Program. Prentice Hall, Upper Saddle (2001); ISBN 0-13-028419-X
5. Turban, E., Lee, J., King, D., Michael Chung, H.: Electronic Commerce A Managerial Perspective. Prentice Hall Inc., Englewood Cliffs (2007); ISBN 0-13-018866
6. Al Shrouf, F.: Facilitator Agent Design Pattern of Procurement Business Systems. In: Proceedings of 32nd Annual IEEE Conference on Computer Software and Applications, COMPSAC 2008, Turku, Finland, pp. 505–510 (2008)
7. Al-Shrouf, F., Turani, A.: Agent Business Systems Development Engineering Approach. European Journal of Scientific Research (EJSR) 29(4), 549–556 (2009); ISSN-1450-216
8. Nwana, H.S.: Software Agents: An Overview. Knowledge Engineering Review 11, 1–40 (1996)
9. Sivan, J.: Building intelligent market places with software agents. Data & Knowledge Engineering 25(1-2), 161–197 (2008)
10. Blake, B.: Innovations in Software Agent based B2B Technologies. In: The 5th International Conference on Autonomous Agents, USA (2001)
11. Blake, B., Conett, T., Piquado, T.: Using Intelligent Agents in Conjunction with B2B Interoperability. In: 1st International Bi-Conference Sessions on Agent-Based Approaches to Business-to-Business (B2B) Interoperability, USA (2001)
12. Gini, M.: Agents and Other Intelligent Software for E-Commerce. In: CSOM (1999)
13. Vetter, M., Pitsch, S.: Using Autonomous Agents to Expand Business Models in Electronic Commerce. Farunhofer Institute for Industrial Engineering, FHG-IAO (2009)

14. Calisti, M.: Intelligent Agents. Artificial Intelligence Lab EPFL (1999)
15. Jennings, N., Sycara, K., Wooldrige, M.: A Roadmap of Agent Research and Development (1998)
16. Pedro, S., Marcus, F., Ayrton, M., Carlos, L.: V-market: A Framework for agent e-commerce system. Software Engineering Laboratory (LES). Computer Science Department, pontifical Catholic University of Rio de Janerio (2004)
17. Gutman, R.H., Moukas, A.G., Maes, P.: Agents As Mediators in Electronic Commerce. Electronic Markets 8, 22–27 (1998)
18. Russell, S., Norvig, P.: Artificial Intelligent A modern Approach. Prentice Hall, Englewood Cliffs (1998)
19. Franklin, S., Graesser, A.: Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents. In: Proceeding of the 3rd International Workshop on Agent Theories (1998)

Mining Opinions in User-Generated Contents to Improve Course Evaluation

Alaa El-Halees

Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
alhalees@iugaza.edu.ps

Abstract. The purpose of this paper is to show how opinion mining may offer an alternative way to improve course evaluation using students' attitudes posted on Internet forums, discussion groups and/or blogs, which are collectively called *user-generated content.* We propose a model to mine knowledge from students' opinions to improve teaching effectiveness in academic institutes. Opinion mining is used to evaluate course quality in two steps: opinion classification and opinion extraction. In opinion classification, machine learning methods have been applied to classify an opinion as positive or negative for each student's posts. Then, we used opinion extraction to extract features, such as teacher, exams and resources, from the user-generated content for a specific course. Then we grouped and assigned orientations for each feature.

Keywords: mining student opinions, E-learning evaluation, opinion mining, student evaluation, opinion classification, opinion extraction.

1 Introduction

The increased use of the Internet has changed people's behavior in the way they express their views and opinions. Nowadays, the quality of products and services are often discussed by customers on the Web. Customers can now post reviews of products and services using Internet forums, discussion groups, and blogs which are collectively called *user-generated content* [1] [2]. In recent years, many researchers used opinion mining to extract knowledge from these user-generated contents. Opinion mining is a research subtopic of data mining aiming to automatically obtain useful knowledge in subjective texts [3]. This technique has been widely used in real-world applications such as e-commerce, business-intelligence, information monitoring and public polls [4].

In this paper we propose a model to extract knowledge from students' opinions to improve teaching effectiveness in academic institutes. One of the major academic goals for any university is to improve teaching quality. That is because many people believe that the university is a business and that the responsibility of any business is to satisfy their customers' needs. In this case university customers are the students. Therefore, it is important to reflect on students' attitudes to improve teaching quality. One way to improve teaching quality is to use traditional student evaluations. The

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 107–115, 2011.

most common criticism of the traditional student evaluations is summarized in three issues; first, it may be biased because students tend to give higher ratings when they expect higher grades in the course [5]. Second, evaluation mainly takes place at the end of a semester. In this case, it is hard to improve the evaluated course and any suggestions will be applied in the subsequent semesters. And, third, regardless of the size of a university, student evaluations generate an enormous quantity of data making the analysis time consuming [6].

To overcome these limitations, we propose to use opinion mining to evaluate course quality. The method can give additional support to traditional student evaluation. There are certain advantages to this method. First, it takes away from traditional classes where most of the time a student takes into account the grade when he/she expresses his/her evaluation. Second, it happens during the semester not at the end, so any recommendations may be taken into account in the same semester not the coming ones. Third, because of the data mining nature which is built for huge data, it is easy to work with an enormous quantity of data generated by students' opinions.

We use opinion mining to evaluate course quality in two steps: opinion classification and opinion extraction. In opinion classification, machine learning methods have been used to classify an opinion as positive or negative for all posts in all courses. Then, in the second step we mined opinions for specific course. In this step, we extracted features for specific courses. Examples of course features contain the teacher and exams. Subsequently, we assigned opinion orientation of the feature (positive or negative). Finally, we grouped the features for each course.

To test our work we collected data from students who expressed their views in discussion forums dedicated for this purpose. The language of the discussion forums is Arabic. As a result, some techniques are used especially for Arabic language.

The rest of the paper is structured as follows: section two discusses related work, section three contains opinion classification, section four is about opinion extraction, section five describes the conducted experiments, section six gives the results of experiments and section seven concludes the paper.

2 Related Work

In publications, we found three works that mentioned the idea of using opinion mining in education. First, Lin et al. in [7] discussed the idea of Affective Computing which they defined as a "Branch of study and development of Artificial Intelligence that deals with the design of systems and devices that can recognize, interpret, and process human emotions". In there work, the authors only discussed the opportunities and challenges of using opinion mining in E-learning as an application of Affective Computing. Second, Song et. al. in [8] proposed a method that uses user's opinion to develop and evaluate E-learning systems. The authors used automatic text analysis to extract the opinions from the Web pages on which users are discussing and evaluating the services. Then, they used automatic sentiment analysis to identify the sentiment of opinions. They showed that opinions extraction is helpful to evaluate and develop E-learning system. Third work of Thomas and Galambos in [9] investigated how students' characteristics and experiences affect their satisfaction. They used regression and decision tree analysis with the CHAID algorithm to analyze student opinion data. They concentrated on student satisfactions such as faculty preparedness, social integration, campus services and campus facilities.

3 Opinion Classification

Opinion mining concerned with enabling system to determine opinions from text written in natural language by human [10]. Opinion classification is a subtopic of opinion mining that classifies an opinionated document as expressing positive or negative. It is also commonly known as sentiment classification or document-level sentiment classification. It aims to find the general sentiment of the author in an opinionated text [4]. For example, in educational data, a student may express his/her opinion about a course using a discussion forum or a blog. Opinion classification determines whether the student attitude is positive or negative about that course.

Formally, given a set of user-generated content review R by students containing opinions about a course, opinion classification aims to classify each document $r \in R$ to determine whether the review is positive, negative or neutral.

We can consider opinion classification as document-level polarity classification which is a special case of text categorization with sentiment positive or negative rather than topic-based categories. With the exception that in traditional document classification, topic words are important. However, with opinion classification, sentiments of the words are more important [11]. Therefore, in opinion classification, we may be able to improve polarity classification by removing objective sentences [10].

In our work, we used modified version of AMOD approach proposed in [12] as follows:

1- A seed set representative of the two adjective categories positive and negative has been provided. Since the work is tested in Arabic language, the seeds are Arabic Adjectives.

2- Synonymy from Online dictionary is used to find new terms that will also be considered representative of the two categories (positive and negative). The new terms, once added to the original ones, yield two new sets of terms.

3- Arabic Datasets, which contained opinion expressions in education, were collected from the Internet.

4- To classify each collected document, we calculate its positive or negative orientation by computing the difference between the number of positive and negative adjectives, from both the previous lists, encountered in the studied document. We count the number of positive adjectives, then the number of negative adjectives, and we simply compute the difference. If the result is positive (greater to given threshold), the document will be classified in the positive class. The same process is done for negative. Otherwise, the document is neutral which is eliminated.

5- We used a binary classifier using the previous documents as training set and user-generated contents as testing set to assign a polarity (e.g. positive or negative) to each student review.

4 Opinion Extraction

Opinion mining discovers opinioned knowledge at different levels such as at clause, feature, sentence or document levels [13]. In the previous section we discussed a way to classify student opinion at document level. This section discusses how to extract opinions in feature level. Features of a product are attributes, components and other aspects of the product. For course improvement feature may be course content, teacher, resources …etc.

We can formulate the problem of extracting features for each course as follows: Given user-generated contents about courses, for each course *C* the mining result is a set of pairs. Each pair is denoted by (*f*, *SO*), where *f* is a feature of the course and *SO* is the semantic orientation of the opinion expressed on feature *f*.

For our work, we used Hu, M., Liu approach in [14]. The goal of the approach is to identify, extract and group the features for each course as follows:

1) Identifying and extracting the course features that have been commented by the students. Traditional Information extraction is used where we search as specific course features in the text.
2) Determining whether the opinion on the feature is positive or negative. In this case we used opinion classification discussed in the previous section.
3) Grouping the features orientations and produce a summary for each course and for each feature in that course.

5 Experiments

To evaluate our method, a set of experiments was designed and conducted. In this section we describe the experiments design including the corpus, the preprocessing stage, the used data mining methods and evaluation metrics.

5.1 Corpus

Initially we collected data for our experiments using 4,957 discussion posts which contain 22 MB of data from three discussion forums dedicated to discuss courses. Then, we focused on the content of five courses including all threads and posts about these courses. Table 1 gives some details about the extracted data. Details of data for each selected course are given in table 2.

Table 1. A summary of the used corpus

Total Number of posts	167
Total Number of Statements	5017
Average number of statements in a post	30
Total Number of Words	27456
Average number of words in a post	164

Table 2. Details about data collected for each of the five courses

Course To Review	Number of Posts	Number of Sentences	Number of Words
Course_1	69	1920	13228
Course_2	34	1321	7280
Course_3	23	617	3587
Course_4	21	524	3183
Course_5	20	635	3407

5.2 Preprocessing

After we collected the data associated with the chosen five courses, we striped out the HTML tags and non-textual contents. Then, we separated the documents into posts and converted each post into a single file. For Arabic scripts, some alphabets have been normalized (e.g. the letters which have more than one form) and some repeated letters have been cancelled (that happens in discussion when the student wants to insist on some words). After that, the sentences are tokenized, stop words removed and Arabic light stemmer applied. We obtained vector representations for the terms from their textual representations by performing TFIDF weight (term frequency–inverse document frequency) which is a well known weight presentation of terms often used in text mining [15]. We also removed some terms with a low frequency of occurrence.

5.3 Methods

In our experiments to classify posts, we applied three machine learning methods, which are Naïve Bays, k-nearest and Support Vector Machine.

Naïve Bays classifiers are widely used because of their simplicity and computational efficiency. It uses training methods consisting of relative-frequency estimation of words in a document as words probabilities and uses these probabilities to assign a category to the document. To estimate the term $P(d \mid c)$ where d is the document and c is the class, Naïve Bayes decomposes it by assuming the features are conditionally independent [16].

k-Nearest Neighbor is a method to classify documents. In the training phase, documents have to be indexed and converted to vector representation. To classify new document d; the similarly of its document vector to each document vector in the training set has to be computed. Then its k nearest neighbor is determined by measuring similarity which may be measured by, for example, the Euclidean distance [17].

Support Vector Machine is a learning algorithm proposed by [18]. In its simplest linear form, it is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. Test documents are classified according to their positions with respect to the hyperplanes.

5.4 Evaluation Metrics

There are various methods to determine effectiveness; however, precision and recall are the most common in this field. Precision is the percentage of predicted reviews class that is correctly classified. Recall is the percentage of the total reviews for the given class that are correctly classified. We also computed the F-measure, a combined metric that takes both precision and recall into consideration [19].

$$F-measure=\frac{2*precision*recall}{precision+recall} \tag{1}$$

6 Experimental Results

We have conducted experiments on students' comments on five selected courses using two steps: Opinion Classification and Opinion extraction.

First, we evaluated opinion classification. Evaluation of opinion classification relies on a comparison of results on the same corpus annotated by humans [20]. Therefore, to evaluate our approach, first we manually assigned a label for each student subjective comment. Then, we used Rapidminer from [21] as data mining tool to classify and evaluate the results of students' posts. Table 3 gives results of the precision, recall and f-measure for each course using three data mining methods Naïve bays, k-nearest and Support Vector Machine. The last row gives the average results.

Table 3. Precision, recall and F-measure of the five courses using three data mining methods

Course	**K-nearest**			**Naïve Bays**			**Support Vector Machine**		
	Pr	**Re**	**F-m**	**Pr**	**Re**	**F-m**	**Pr**	**Re**	**F-m**
Course_1	69.23	75	72	76.7	66.7	71.35	70	72.16	71.06
Course_2	79.09	59.09	72.36	84.29	81.82	83.04	74.74	81.21	77.84
Course_3	58.24	72.23	64.48	67.67	93.94	78.69	61.25	88.15	72.28
Course_4	70	53.85	60.87	72.5	76.92	74.65	69.09	75.52	72.02
Course_5	87.5	82.35	84.85	86.74	76.74	81.43	80	94.12	86.49
Average	70.81	68.50	70.91	77.58	79.22	77.83	71.02	82.23	76.22

From the table, with precision of 77.58 %, we can conclude that Naïve Bays method has better performance than the other two methods. However, with recall of 82.23% Support Vector machine has better performance. Overall, Naïve Bays has the best f-measure with 77.83%.

In the second step, we selected a set of features for course evaluation. From the data, we found that the most frequent features are Teacher, Content, Exams, Marks and Books. We used Gate Information Extraction tool from [22] for feature extractions .Then, we used the system assignment proposed in opinion classification to assign the orientation of the posts. After that, we grouped the features. Figure 1 gives an

Course_1:
Contain:
Positive: 18
Negative: 25
Teacher:
Positive: 28
Negative: 19
Exams:
Positive: 18
Negative: 21
Marks:
Positive: 20
Negative: 6
Books:
Positive: 11
Negative: 21

Fig. 1. Feature_ based opinion extraction for course_1

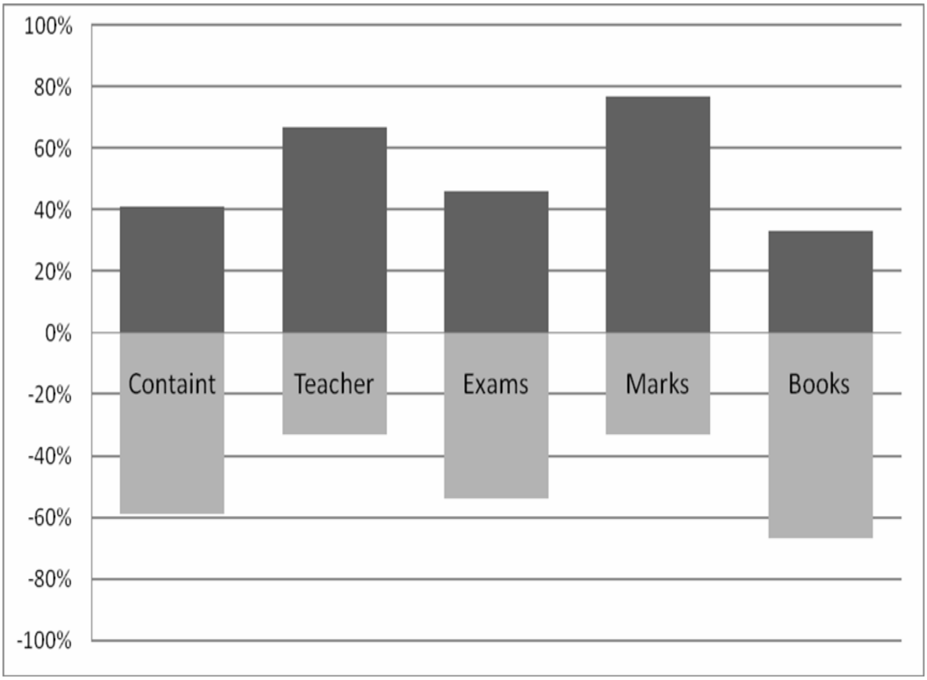


Fig. 2. Graph of feature_ based opinion extraction for course_1

example of features extraction for course_1. Figure 2 visualizes the opinion extraction summary as graph.

In figure 2, it is easy to envisage the positive and negative opinions for each feature. For example, we can figure out that *Books category* has negative attitude while *marks category* has positive attitude from the point of view of the students.

7 Conclusion

The aim of this work is to present the usefulness of discovering knowledge from user-*generated content* to improve course performance. Our goal is to supplement student evaluation of courses, not to replace the traditional way of course evaluation. We used opinion mining in two steps: first to classify student posts for courses where we used three machine learning methods. Then, to extract and group features for each course we used opinion extraction method.

We think this is a promising way of improving course quality. However, two drawbacks should be taken into consideration when using opinion mining methods in this case. First, if the student knew that his posts will be used for evaluation, then he/she will behave in the same way of filling traditional student evaluation forms and no additional knowledge can be found. Second, some students, or even teachers may put spam comment to bias the evaluation. However, for latter problem methods of spam detection, such as work of [23], can be used in future work. Also, comparing courses for each lecturer or semester could be useful for course evaluations.

References

[1] Liu, B.: Searching Opinions in User-Generated Contents. In: Invited talk at the Sixth Annual Emerging Information Technology Conference (EITC 2006), Dallas, Texas, August 10-12 (2006)

[2] Leung, C.W.K., Chan, S.C.F.: Sentiment Analysis of Product Reviews. In: Wang, J. (ed.) Encyclopedia of Data Warehousing and Mining Information Science Reference, 2nd edn., pp. 1794–1799 (August 2008)

[3] Song, H., Yao, T.: Active Learning Based Corpus Annotation. In: IPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, pp. 28–29 (August 2010)

[4] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Information Retrieval 2, 121–135 (2008)

[5] Huemer, M.: Student Evaluations: a Critical Review, `http://home.sprynet.com/~owl/.sef.htm` (accessed in January 2011)

[6] Kozub, R.M.: Student Evaluations Of Faculty: Concerns And Possible Solutions. Journal of College Teaching & Learning 5(11) (November 2008)

[7] Lin, H., Pan, F., Wang, Y., Lv, S., Sun, S.: Affective Computing in E-learning. E-learning, Marina Buzzi, InTech, Publishing (February 2010)

[8] Song, D., Lin, H., Yang, Z.: Opinion Mining in e-Learning. In: IFIP International Conference on Network and Parallel Computing Workshops (2007)

[9] Thomas, E.H., Galambos, N.: What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis. Research in Higher Education 45(3), 251–269 (2004)

[10] Xia, L., Gentile, A.L., Munro, J., Iria, J.: Improving Patient Opinion Mining through Multi-step Classification. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 70–76. Springer, Heidelberg (2009)

[11] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)

[12] Harb, A., Plantié, M., Dray, G.: Web opinion mining: how to extract opinions from blogs? In: Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology. ACM, New York (2008)

[13] Balahur, A., Montoyo, A.: A Feature Dependent Method For Opinion Mining and Classification. In: International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE, pp. 1–7 (2008)

[14] Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA (2004)

[15] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)

[16] Du, R., Safavi-Naini, R., Susilo, W.: Web filtering using text classification (2003), http://ro.uow.edu.au/infopapers/166

[17] Dasarathy, B.: Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos (1991)

[18] Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20 (1995)

[19] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA (February 1999)

[20] Osman, D., Yearwood, J.: Opinion search in web logs. In: Proceedings of the Eighteenth Conference on Australasian Database, Ballarat, Victoria, Australia, vol. 63 (2007)

[21] http://www.Rapidi.com

[22] http://gate.ac.uk/

[23] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H.: Detecting Product Review Spammers using Rating Behaviors. In: The 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, October 26 - 30 (2010)

Rough Set Theory Approach for Classifying Multimedia Data

M. Nordin A. Rahman, Yuzarimi M. Lazim, Farham Mohamed,
Suhailan Safei, Sufian Mat Deris, and M. Kamir Yusof

Faculty of Informatics, Universiti Sultan Zainal Abidin,
Gong Badak Campus 21300 Kuala Terengganu, Malaysia
mohdnabd@unisza.edu.my, ayulazim@yahoo.com.my

Abstract. The huge size of multimedia data requires for efficient data classification and organization in providing effective multimedia data manipulation. Those valuable data must be captured and stored for potential purposes. One of the main problems in Multimedia Information System (MIS) is the management of multimedia data. As a consequence, multimedia data management has emerged as an important research area for querying, retrieving, inserting and updating of these vast multimedia data. This research considers the rough set theory technique to organize and categorize the multimedia data. Rough set theory method is useful for exploring multimedia data and simplicity to construct multimedia data classification. Classification will help to improve the performance of multimedia data retrieving and organizing process.

Keywords: Rough set theory, multimedia data management, approximation, classification, data clustering.

1 Introduction

Everyone deals with multimedia data at every walk of lives. Multimedia data consist of texts, graphics, animations, video, sounds, music etc. People are working with multimedia data and surrounded by them. Therefore, there are many issues and challenges faced by multimedia data providers to fulfill the user requirements. One of the issues is to organize and classify the huge multimedia data so that the information can be obtained easily at any point of time. An efficient multimedia data management is highly required because it will improve the process of multimedia information discovery especially for decision making application, business marketing, intelligent system, etc [1]. To do so, multimedia data management is a tool required to manage and maintain huge multimedia data.

Rough set theory is an effective tool for classification applications and, it has been introduced by Pawlak [2] [3]. Rough set theory is a mathematical tool that can be used for processing and analyzing of inexact, uncertain, and vague datasets. It is an extension of set theory for study of the intelligent system characterized by insufficient and incomplete information [3]. Various efforts have been made to improve the efficiency and effectiveness of classification with rough sets [4]. Practically, rough set theory has been applied to the number of application domains such as medical

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 116–124, 2011.

diagnosis, engineering reliability, expert systems, empirical study of materials data, machine diagnosis, business failure prediction, activity-based travel modeling, travel demand analysis, solving linear programming and data mining [5].

The intention of this research is to introduce a new representation for multimedia data management as an information system by integrating rough set theory elements such as decision table and approximation. This paper is organized as follows: Section 2 describes the related work on all the issue in multimedia data management while Section 3 gives details explanation on basic concepts of rough set theory. Section 4 demonstrates the proposed framework and Section 5 draws the conclusion.

2 Multimedia Data Management

Multimedia is defined as combination of more than one media; they may be of two types, static and dynamic media. Text, graphics and images are categorized as static media, while objects like animation, music, audio, speech, and video are categorized as dynamic media [6]. Multimedia data contains an enormous amount of information. This information is in the form of identifiable "features" in the multimedia data. For example, video data contains timing data that can be used to track the movement of an object from frame to frame or to identify transitions between scenes. Similarly, audio data contains certain identifiable features such as words, sound, pitches, and silent periods as well as timing information [7].

Multimedia data is attractive, informative and stored in compact format. It has been used by various categories of user; from technical up to management levels. The growing of consumer demand for multimedia information makes sophisticated technology is needed in multimedia data management. The applications are including representing, modeling, indexing, retrieving and manipulating. The explosion of multimedia content in various aspects such as in databases, broadcast, steaming media, etc. has generated new requirements for more effective access to these global information repositories. Multimedia data requires for a huge storage area and each media type requires different methods to store and retrieve.

The major issues that related to multimedia data management system are multimedia data modeling, huge capacity storage management, information retrieval capabilities, media integration and presentation. Multimedia database management system (MDMS) is developing in purpose to fulfill this requirement. MDMS supports facilities for the indexing, storage, retrieval and provides a suitable environment for using and managing multimedia data [8]. Technique of indexing and classification in multimedia data is created in order to ease the query processing.

Another study [9] designed a software tool known as web-based multimedia news management system (MNMS) which consist of a collection system, website management system, workflow system and publishing system. The MNMS tool enables delivery of TV and newspaper content trough ITV with internet connection. The integrated both publications are identify as multimedia interactive digital news. MNMS provides a service and application to allow collaboration and communication among journalists, reporters, editors and designers from multinational publication company that distributed around the world. The concept of MNMS has been implemented as part of the data broadcasting in the project entitled News On Demand Kiosk Network (NODKN) [9]. This project is a collaborative between Multimedia University in Malaysia and Mitsushita in Japan.

The most essential activity involves in news content management process are authoring and publishing with additional of multimedia elements such as video, audio, image, text and animation. The MNMS tool provides an environment where the employees of a publication company can create, reviewing, publishing, versioning, archiving, changing and delete their news content or items. They can use this tool to support authoring and publishing operation of multimedia interactive news. The MNMS also provide a medium to communication with other user around the world to ease them shared the information. As a web application, the MNMS system allows end-users to search, view and read articles the interactive multimedia news based on their own personalization; for instance they can choose their own language.

Large volume patient related information like demographic data, historical data, data-rich multimedia diagnostic studies and longitudinal disease are one of the important parts in medicine information management. Some of general medicine clinics still used traditional record system such as handwritten notes to record information from patient examination and the physician just can review the record after accessing it in the office. To overcome this situation, Electronic Medical Record (EMR) application was suggested by [10] in order to manage patient data efficiently and accurately.

The EMR tool was develop in Spanish and English version language with the clinical practitioner at remote clinics in Ecuador. The tools support for patient information management in electronic form and electronic file sharing between the regional collaborating clinics in Ecuador. The implementation of EMR, migrates the patient information management from handwriting notes to distributed health care system. An authorize staff or clinicians easily recorded patient information and were able to achieve the information back for updating or for follow-up in patient care practices. Images or video data captured by radiographs, sonograms, microscopic and colposcopy examinations process were composed and stored within the workstation. The EMR user friendly interface allowed quick load and review of the digital images when the data is needed.

An unstructured data such as multimedia files are difficult to capture and store in the common database storage. In one of such study, [11] was interested to design a prototype tool to extract and classify unstructured data in any web pages. The prototype was focus on an efficient data classification and organization in providing effective multimedia data manipulation. Document Object Module (DOM) tree technique is applied to find the correct data in the HTML document in classification process. The prototype was implemented and tested on four main class data type in various formats to help end user get useful multimedia data stored for future retrieval usage. However, some of unnecessary information, such as script, style or other customized nodes in DOM tree cannot eliminate completely and may be minimized during extraction process.

Several weaknesses in current multimedia data management model have been found, and the weaknesses are as follow:

- Various models do not combine all type of multimedia data for classification but focusing on one type of data in each research
- Previous studies have emphasized that the issue of multimedia data storage and management, but little is so far known about the classification of multimedia data
- A lot of multimedia data classification model based on media format (e.g : .jpg, .txt, .mp3, .flv) but not by their attribute.

3 Rough Set Theory

Rough set theory, introduced by Pawlak [2][3] in the early 1980s. It is an extension of the set theory for the study of intelligent system characterized by inexact, uncertain or vague information and can serve as a new mathematical tool to soft computing [12]. An upper and a lower approximation of a set, the approximation space and models of sets are the fundamental concepts in rough set theory [13]. General elements engage in rough sets theory can be described as follows:

1) *Universe:* A non-empty finite of objects named training set, $U = \{x_1, x_2, \ldots, x_n\}$;
2) *Attributes:* A non-empty finite set of attributes, $A = \{a_1, a_2, \ldots, a_k\}$;
3) *Information system:* (also called decision table) is a pair of the universe and attributes, IS=<U,A> ;
4) *Indiscernibility relation:* (known as equivalence class) defines a partition in U. The indiscernibility relation is defined as, $R(B) = \{(x,y) \in U \times U : \text{for all } a \in B, a(x) = a(y)\}$ where, $a \in A$ and $B \subseteq A$;
5) *Approximation spaces:* define as a pair, AS = <U, R(C)> where, C be a set of condition attributes and R(C) be an indiscernibility relation on U.

   - $[x]_B$ denotes the equivalence class of B containing x, for any element x of U;
   - Based on singleton x, for a given $B \subseteq A$ and $X \subseteq U$, the *lower approximation* ($\underline{B}X$) of the set X in IS and the *upper approximation* of the set X in IS ($\overline{B}X$) are defined as follows:

$$\underline{B}X = \{x \in U : [x]_B \subseteq X\}. \quad (1)$$

$$\overline{B}X = \{x \in U : [x]_B \cap X \neq \varnothing\}. \quad (2)$$

   - For a given $B \subseteq A$ and $X \subseteq U$, the boundary of X in IS can be defined as,

$$\mathrm{BND}(X) = \overline{B}X - \underline{B}X. \quad (3)$$

BND(X) consists of objects that do not certainly belong to X on the basis of A.

Data used in rough set theory are often presented as a table which is initialized as decision table as illustrated in figure 1. In the table, columns correspond to attributes and rows of the decision table correspond to objects. Entries in the table are attribute values. The attributes of the decision table include condition attribute and decision attribute. The attributes in A can be further classified into two disjoint subsets, *condition attributes* (C) and *decision attributes* (D) such that A=C $\cup$ D and C $\cap$ D = $\varnothing$. The decision attributes can have some values though quite often it is binary [14]. Let B $\subseteq$ A, U/R (B) denotes a family of all equivalence classes of the relation R (B) called elementary sets. The elements of $\underline{B}$ X are all and only those objects x $\in$ U which belong to the equivalence class generated by the indiscernibility relation contain in X. Meanwhile, the elements of $\overline{B}$ X are all those objects x $\in$ U which belong to the equivalence classes generated by the indiscernibility relation containing at least on object x belong to X. The BND (X) indicates the objects in IS are inconsistent or vague. If upper and lower approximations of X are equal then X is an ordinary set. Clearly that, rough set theory mainly resolves to the problem how X $\subseteq$ U can be c-overed with a set of equivalence classes according to indiscernibility relation.

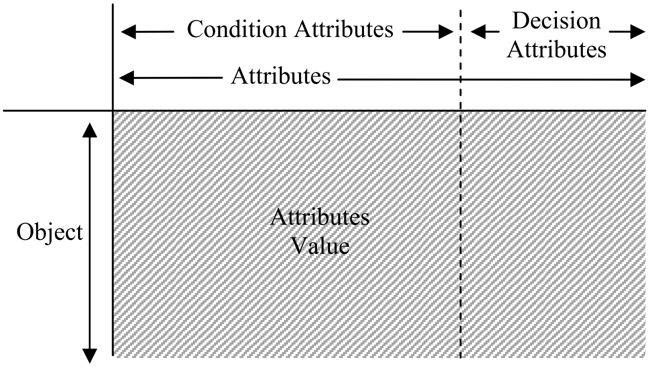


Fig. 1. Decision table

4 Classification Process for Multimedia Data Using Rough Set Theory

Classification of objects in the databases or information systems sources based on rough set theory has been done in many applications [5][15][16]. The goal of classification is to build a set of models that can accurately predict the class of different objects. Rough set mainly deals with data analysis in table format. The approach is generally to process the data in the table and then to analyze them. In this section, several multimedia data sets are shown with possible media type will be used as an example to illustrate the concept of rough set theory. Let, an information system is a set of objects represented in a data table, the rows are considered as objects for analysis and the columns represent a measureable attributes for each object, where

IS = <U,A>. Table 1 shows an information system which is a collection of multimedia data as object.

Given a set of universe, U= {O_1, O_2, O_3, O_4, O_5, O_6, O_7}, where U are all of the objects. The set condition attributes is represented by A= {*Illustration, Timeline, Movement*} and the set D represented the decision attribute, where D= {Media Types}. Table 1 can be shown in relation to the function of nominal values of the considered attributes, in the Table 2. Based on Table 1, classification of multimedia data produced based on condition attribute for each object. Theoretically, video and audio contain timing data [7] which can be used to track the movement of an object from frame to frame. Images are categorized as static media, while audio and video are categorized as dynamic media [6]. To build this information system, if C = {*Illustration*, yes} then, decision attribute can be defined as video or image media types. If C = {*Timeline*, yes} then decision attribute as video or audio types are admitted certainly. Justification based on attribute can be used to classify media types whether; it is a video, an audio or an image.

Table 1. Information system

	Condition Attributes			**Decision Attribute**
Object	*Illustration*	*Timeline*	*Movement*	Media Types
O_1	Yes	Yes	Dynamic	Video
O_2	Yes	No	Static	Image
O_3	Yes	No	Dynamic	Video
O_4	No	Yes	Static	Audio
O_5	Yes	Yes	Static	Video
O_6	No	Yes	Dynamic	Audio
O_7	No	Yes	Dynamic	Video

Table 2. Nominal Values of Attributes

	Attributes	**Nominal Values**
Condition Attributes	*Illustration*	Yes, No
	Timeline	Yes, No
	Movement	Static, Dynamic
Decision Attributes	Media Types	Video, Audio, Image

Indiscernibility relation is the relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. In Table 1, it can be observed that the set is composed of attributes that are directly related to multimedia data, where A={*Illustration, Timeline, Movement*}, the indiscernibility relation is given to R(A). When Table 1 is broken down it can be seen that the set regarding {O_1, O_2, O_3, O_5} is indiscernible in terms of *Illustration* attribute. The set concerning

{O_1, O_4, O_5, O_6, O_7} is indiscernible in term of *Timeline* attribute, and the *Movement* attribute generates two indiscernibility elementary sets are {O_1, O_3, O_6, O_7} and {O_2, O_4, O_5}.

Approximations are fundamental concepts of rough set theory, it is can be defined as upper bounds and lower bounds. As define in section 3, based on objects in Table 1 the lower and upper approximation of X are categorize as follows,

$\underline{B}(X) = \{O_1,O_2,O_3,O_4,O_5\}$, $\overline{B}(X)=\{O_1,O_2,O_3,O_4,O_5O_6,O_7\}$. As a result,

BND(X) = {O_6, O_7}. Figure 2 shows the relationship of the lower and upper approximation of an information system. The elements that enclosed by thick line belong to upper approximation. Meanwhile, the elements that enclosed by light line belong to the original set X. The elements that covered by grey color is belong to the lower approximation.

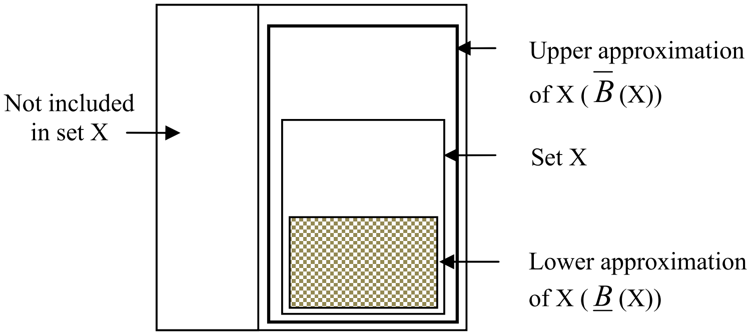


Fig. 2. Lower and upper approximation of set X

Table 3. Reducts table

	Condition Attributes		**Decision Attribute**
Object	*Illustration*	*Timeline*	Media Types
$O_{1\,,}O_5$	Yes	Yes	Video
O_2	Yes	No	Image
O_3	Yes	No	Video
$O_{4,}O_6$	No	Yes	Audio
O_7	No	Yes	Video

Based on lower and upper approximation, one way to facilitate data retrieving and manipulation is by reduction the set of data with reducing attributes. In indiscernibility relation, only attribute that do not contribute to the classification result can be omitted. Reduction means, the set of remaining attributes is the minimal set, and set which presents in all subsets call cores, in other words, removing repetitive or overlapping data. The main purpose of reduction is to determine the attributes which can represent data in a database and dependencies between attributes. Table 3 shows the example of reduct which drops attribute *movement* and combined same objects in the same row. Decision rules 4) and 5) in table 3 have the same conditions but different decisions. Such rules are called inconsistent; otherwise the rules are referred to as consistent.

Decision rule created by combining rule reducts attributes. Each rows of reduct table verify a decision rule, which specifies the decision that must be taken when condition are indicated by condition attributes are fulfilled. Decision rules frequently presented as implication called "*if...then...*" rules. From the certainty factors of decision rules, the result as below:

- *if (illustration,yes) and (timeline,no) then (media types,image)*
- *if (illustration,no) and (timeline,yes) then (media types,audio)*
- *if (illustration,yes) and (timeline,yes) then (media types,video)*

Classifying data into several attributes is important because the attribute has to be matched with the corresponding data classes specified in the decision attribute. By applying rough set theory in classification of multimedia data, this model consists of six important elements; information system, indiscernibility relation, lower and upper approximation, reduction and decision rules. This model demonstrates that redefined indiscernibility can reduce the number of elementary sets. In addition sets of one object will enhance the approximation precision, but decrease the accuracy of decisions. Using the model provided, the theory of rough set proves to be an effective tool for multimedia data because:

- It reduces the data set without losing originality of characteristic set
- It is easy for clustering data
- It can manage multimedia data and effortless for application access.

5 Conclusion

The proposed model has been developed to help the end users to manage useful multimedia data (audio, video and image) in order to allow efficient process for storing, retrieving and updating data. The fundamental concept for classifying multimedia data in this research is based on attributes; *illustration, timeline* and *movement*. These attributes are mainly used to make a decision in classifying of media types. This research also has a new contribution in introducing rough set theory technique to organize and categorize multimedia data. The integration of classification data using rough set theory is believed to improve multimedia data management process. With more comprehensive study and investigation, we assume that some applications using classification of attribute in the rough set theory, using existing multimedia data in the real life multimedia organization through this view will be applicable. A future vision is to investigate the performance of this proposed model executed under the web services.

Acknowledgements

This research was supported by Fundamental Research Grants Scheme 2009 (FRGS-2009), Ministry of Higher Education, Malaysia and Universiti Sultan Zainal Abidin.

References

1. Nordin, M.A.R., Farham, M., Suhailan, S., Sufian, M.D., Kamir, M.Y.: Applying Time Granularity in Multimedia Data Management. In: Proc. International Conference on Computational Intelligence and Vehicular System (CIVS), pp. 60–64 (2010)
2. Pawlak, Z., Grzymala–Busse, J.W., Slowiriski, R., Ziarko, W.: Rough Sets. Comm. of the ACM. 38(11), 88–95 (1995)
3. Pawlak, Z.: Rough Set: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrent (1991)
4. Zhong, N., Dong, J.Z., Ohsuga, S.: Using Rough Sets with Heuristics for feature Selection. Journal of Intelligent Information Systems 16, 199–214 (2001)
5. Shyng, J.-Y., Wang, F.-K., Tzeng, G.-H., Wu, K.-S.: Rough Set Theory in Analyzing the Attributes of Combination Values for the Insurance Market. Journal of Expert Systems with Application 32, 56–64 (2007)
6. Jalal, S.K.: Mutimedia Database: Content and Structure. In: Workshop on Multimedia and Internet Technologies, Bangalore (2001)
7. Griffioen, J., Seales, B., Yavatkar, R., Kiernan, K.S.: Content Based Multimedia Data Management and Efficient Remote Access. Extrait de la Revue Informatique et Statistique dens les Sciences Humaines 1(4), 213–233 (1997)
8. Candan, K.S., Sapino, M.L.: Data Management for Multimedia Retrieval. Cambridge University Press, New York (2010)
9. Cheong, S.N., Azahar, K.M., Hanmandlu, M.: Development of Web-based Multimedia News Management System for News on Demand Kiosk Network. WSEAS Transaction on Computers 2(2), 360–365 (2003)
10. Azhar, R., Zhao, X., Cone, S., Merrell, R.: Electronic Multimedia Data Management for Remote Population in Ecuador. International Congress Series 1268, 301–306 (2004)
11. Abidin, S.Z.Z., Idris, N.M., Husain, A.H.: Extraction and Classification of Unstructured Data in WebPages for Structured Multimedia Database via XML. In: International Conference in Information Retrieval Knowledge Management (CAMP), pp. 44–49 (2010)
12. Xu, W.-H., Zhang, W.-X.: Knowledge Reduction in Consistent Information System Based on Dominance Relations. In: Liu, Y., Chen, G., Ying, M. (eds.) Optimization Techniques 1973. LNCS, vol. 3, pp. 1493–1496. Springer, Tsinghua University Press (2006)
13. Wang, Y., Ding, M., Zhou, C., Zhang, T.: A Hybrid Method for Relevance Feedback in Image Retrieval Using Rough Sets and Neural Networks. International Journal of Computational Cognition 3(1), 78–87 (2005)
14. Fomina, M., Kulikov, A., Vagin, V.: The Development of The Generalization Algorithm based on The Rough Set Theory. International Journal Information Theories & Application 13(13), 255–262 (2006)
15. Hu, X.: Using Rough Sets Theory and Database Operation to Construct a Good Ensemble of Classifiers for Data Mining Applications. In: Proc. of ICDM, pp. 233–240 (2001)
16. Nordin, M.A.R., Yazid, M.M.S., Aziz, A., Osman, A.M.T.: DNA Sequence Database Classification and Reduction: Rough Sets Theory Approach. In: Proc. of 2nd International Conference on Informatics, pp. 41–47 (2007)

Application of Rasch Model in Validating the Content of Measurement Instrument for Blog Quality

Zuhaira Muhammad Zain[1], Abdul Azim Abd Ghani[1], Rusli Abdullah[1],
Rodziah Atan[1], and Razali Yaakob[2]

[1] Department of Information Science,
Universiti Putra Malaysia
Serdang, Selangor
zuhaira.muhdzain@gmail.com
[2] Department of Computer Science,
Universiti Putra Malaysia,
Serdang, Selangor
{azim,rusli,rodziah,razaliy}@fsktm.upm.edu.my

Abstract. Research in blog quality is very crucial nowadays in order to have a good quality blog in the blogosphere. The blog quality criteria have been derived from a rigorous metadata analysis. Yet, these criteria have not been reviewed and their significance has not been proven systematically. In this paper, Rasch Model is applied to produce an empirical evidence of content validity of the blog quality criteria. This study confirms that the definitions of 11 families and the 49 criteria assigned have content validity by mean of online survey. These criteria will then be used as a basis of constructing the instrument to measure the acceptability of the criteria for blog quality.

Keywords: Blog quality, content validity, Rasch Model.

1 Introduction

Advances in technology are making use of Internet as an ever-growing phenomenon and we are witnessing a tremendous growth of blogs in the blogosphere. As reported by Pew Internet & American Life Project Surveys in 2006, there were 12 million of American adults who keep a blog [1]. We believe that the current figure has risen, due to the increased of broadband penetration rate that has exposed all levels of users to blogging. As stated in the World Broadband Statistics Report, by the end of 2009, there were 466.95 million broadband subscribers worldwide, up to 2.5 per cent on the previous quarter from 455.57 million [2]. In addition, the technical skills required to create a blog are readily available. The emergence of user-friendly and free blog tools such as Blogger, WordPress, and LiveJournal simplifies the process of web publishing, which formerly required some programming skills [3]. Blogs may offer vital information for blog readers to make decisions or keep abreast with the latest blog posting. Yet, the growth presents disorganized and uncontrolled, thus contributing to the limitation of having bad blogs in the blogosphere. Consequently, blog readers

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 125–136, 2011.

have access to bad or poor-quality blogs that may turn off readers' interest to return to the same blog. Fulfilling readers' expectations and needs is necessary for developing good quality blog.

This study endeavoured into the criteria required in blog quality that will promote to readers satisfaction. The blog quality criteria have been derived from a thorough metadata analysis. The criteria have been consolidated from literatures and researches done in the area of blog, website design, information quality on the Web and portal data quality. However, these criteria have not been reviewed and their content validity has not been verified systematically.

Content validity test is a subjective assessment of how suitable an instrument is to a group of reviewers. It involves a systematic review of the survey's contents to ensure that it includes everything it should, and excludes everything it should not. It is also very important in providing a good foundation on which to base a rigorous assessment of validity [4]. However, Kitchenham and Pfleeger [4] argued that there is no content validity statistic. Later, this argument has been refuted by Abdul Aziz et al. [5] where they proved that content validity can be assessed accurately and fast despite the small sample size by using Rasch Measurement Model. It is a measurement model that is formed as a result of the consideration that takes into account the ability of the respondents, and the difficulty of items [6]. The graphical output provided is great which gives better clarity for quick and easy decision making [7].

The purpose of this study is to determine content validity of the measurement instrument for blog quality. It is to confirm whether the prescribed family definitions and criteria assigned are agreeable to the reviewers. This study proves that the definitions of 11 families and the 49 criteria assigned have content validity by mean of online survey.

The rest of this paper is organized as follows. Section 2 describes the basic of Rasch measurement method. Section 3 explains how content validity test is conducted. Section 4 discusses the results of this study. Finally, Section 5 touches on the conclusions and future work.

2 Rasch Measurement Method

Response from the experts on the content validity is considered rating scale in which the experts rated the criteria according to their agreement. In theory, at this phase, the study is only counting the number of positive answers from the experts which is then added up to give a total raw score. The raw score only provides a ranking order which is supposed an ordinal scale that is continuum in nature [8]. It does not have equal intervals which contradicts the nature of numbers for statistical analysis and it does not meet the fundamentals of adequate statistics for evaluation [9].

Rather than fitting the data to suit a measurement model with errors, Rasch focuses on developing the measurement instrument with accuracy. By emphasizing on the reproducibility of the latent trait measurement instead of forcing the expected generation of the same raw score, i.e. the common expectation on repeatability of results being a reliable test, the concept of reliability takes its rightful place in supporting validity rather than being in contentions. Consequently, measuring quality in an appropriate way is vital to ensure valid quality information can be produced for meaningful use; by absorbing the error and representing a more accurate prediction based on a probabilistic model [10].

In Rasch measurement Model, the probability of success can be estimated for the maximum likelihood of an event as;

$$P(\theta) = \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}} \tag{1}$$

where;

e = base of natural logarithm or Euler's number; 2.7183
β_n = person's ability
δ_i = item or task difficulty

3 Methodology

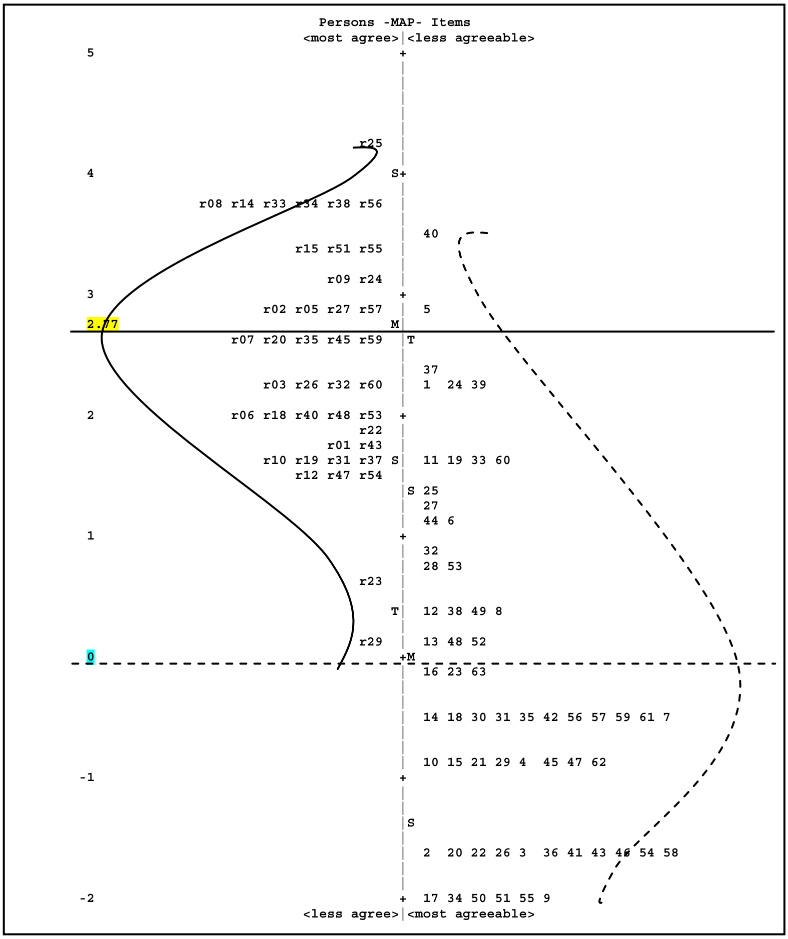
Experts from the field of English and Information Technology who read blogs (comprising 50 English Language lecturers from various universities in Malaysia and 50 Information Technology executives or managers from various companies of more than 10 years working experience) form the pool of reviewers. Therefore, the total number of experts engaged at 100.

The descriptive design chosen is based on the survey objective. It is suitable for checking whether the experts agree with the proposed set of definition of family and the assigning of criteria to the respective families. The questionnaire consists of closed questions (with a Yes/ No answer being provided) and also an open question for differing views and comments. In this context, the endorsement of the experts is of interest.

Invitation by e-mail to join the online survey is being sent to the subjects. The objective of the study; its relevance; the importance of individual's participation; and the confidentiality assured, have been made known for the purpose of the survey. The feedback was then tabulated and analyzed using Rasch Measurement Model [11] with the aid of Rasch analysis software [12].

4 Results and Discussions

A total of 60 participants out of 100 subjects submitted the online survey forms (30 respondents being experts in English with another 30 from Information Technology). This represents a response rate of 60 percent. The responses are then tabulated and run in Bond&FoxSteps software in order to obtain the *logit* measures.



```
                         Persons -MAP- Items
                      <most agree>|<less agreeable>
   5                              +
                                  |
                                  |
                                  |
                                  |
                                  |
                              r25 |
   4                             S+
                                  |
         r08 r14 r33 r34 r38 r56  |
                                  |
                                  |  40
                     r15 r51 r55  |
                                  |
                         r09 r24  |
   3                              +
                 r02 r05 r27 r57  |  5
2.77                             M|
            r07 r20 r35 r45 r59   |T
                                  |
                                  |  37
                 r03 r26 r32 r60  |  1  24 39
                                  |
   2        r06 r18 r40 r48 r53   +
                             r22  |
                         r01 r43  |
                 r10 r19 r31 r37 S|  11 19 33 60
                     r12 r47 r54  |
                                  |S 25
                                  |  27
                                  |  44 6
   1                              +
                                  |  32
                                  |  28 53
                             r23  |
                                  |
                                 T|  12 38 49 8
                                  |
                             r29  |  13 48 52
   0                             +M
                                  |  16 23 63
                                  |
                                  |
                                  |  14 18 30 31 35 42 56 57 59 61 7
                                  |
                                  |
                                  |  10 15 21 29 4  45 47 62
  -1                              +
                                  |
                                  |
                                  |S
                                  |
                                  |
                                  |  2  20 22 26 3  36 41 43 46 54 58
                                  |
                                  |
  -2                              +  17 34 50 51 55 9
                       <less agree>|<most agreeable>
```

Fig. 1. PIDM

The Person-Item Distribution Map (PIDM) as shown in Fig. 1 portrays experts' agreement and the acceptability of the 11 family definitions with the 52 criteria assigned. Acceptability Test Level of Agreement given by experts for the identified family and blog quality criteria can be easily established by using formula in Equation 1:

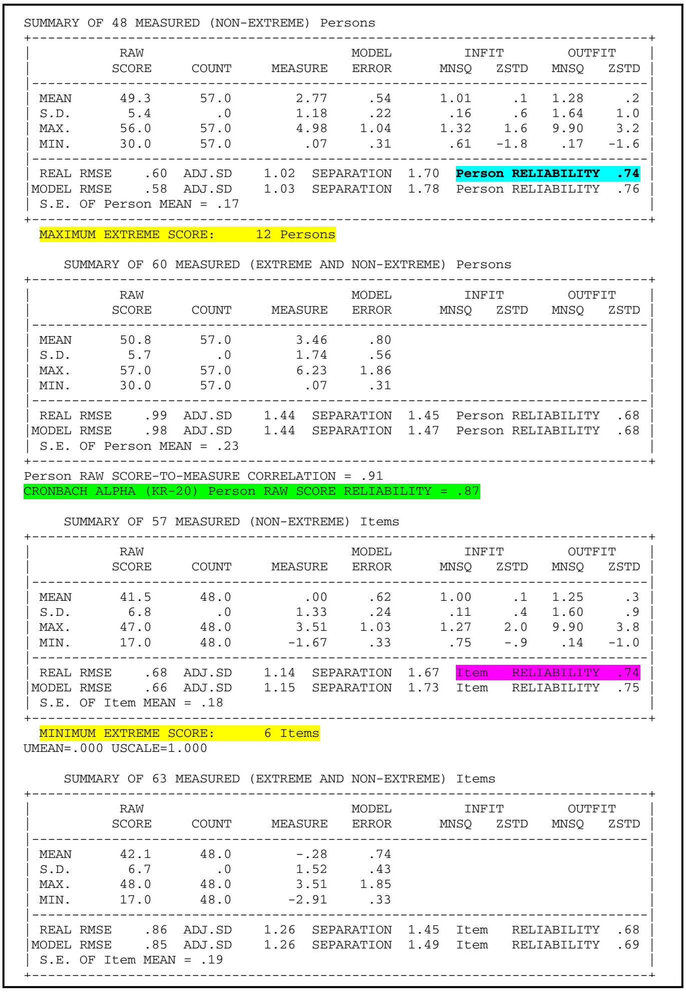
$$P(\theta) = \frac{e^{2.77-0}}{1+e^{2.77-0}}$$

$$= 0.941$$

Thus, experts have indicated their Level of Agreement at 94.1% which is above the 70% threshold limit of Cronbach Alpha. Hence, all experts agree to the prescribed family definition and criteria assigned. This can be determined clearly from PIDM where the person mean μ_{person}=+2.77 *logit* is located higher than the item mean; μ_{item}

which is constrained to 0.00 *logit*. This indicates that all experts involved in the Content Validity Test have the tendency for agreeing to the entire family definitions and criteria assigned that have been proposed.

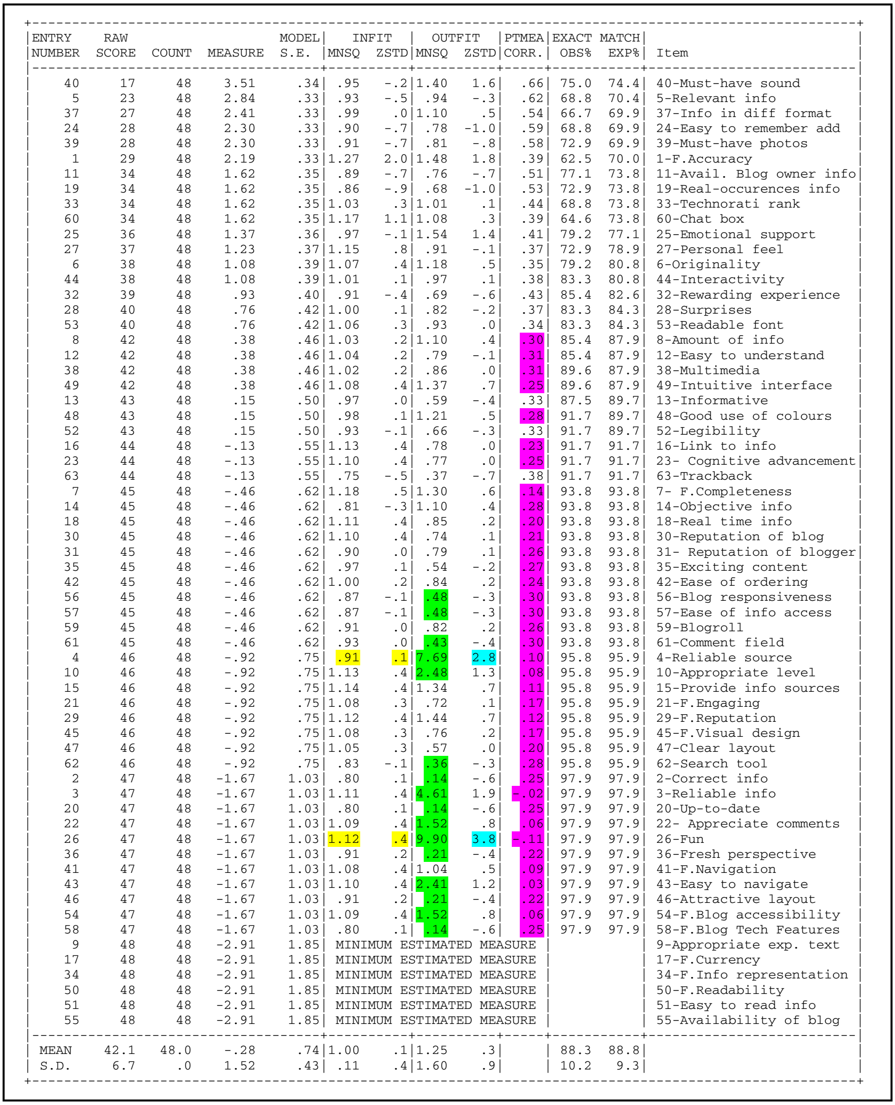
The Summary Statistics in Fig. 2 shows three important indicators; Cronbach Alpha value, Person Reliability value, and Item Reliability value. The summary shows that there is a good value of Cronbach Alpha (0.87). In addition, there is a fair value of Person Reliability (0.74) and Item Reliability (0.74). The values of the three indicators (> 0.6) do confirm that the instrument for measuring content validity is reliable, reproducible, and valid for measurement. The summary also depicts that there are 12 persons and 6 items with maximum and minimum extreme score respectively. This means that there are 12 experts who agreed with all the definitions and criteria assigned. Also we have 6 items which are 100% agreed by the experts.



```
SUMMARY OF 48 MEASURED (NON-EXTREME) Persons
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       49.3      57.0        2.77     .54      1.01     .1   1.28     .2 |
| S.D.        5.4        .0        1.18     .22       .16     .6   1.64    1.0 |
| MAX.       56.0      57.0        4.98    1.04      1.32    1.6   9.90    3.2 |
| MIN.       30.0      57.0         .07     .31       .61   -1.8    .17   -1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .60  ADJ.SD    1.02  SEPARATION  1.70  Person RELIABILITY  .74 |
|MODEL RMSE    .58  ADJ.SD    1.03  SEPARATION  1.78  Person RELIABILITY  .76 |
| S.E. OF Person MEAN = .17                                                   |
+-----------------------------------------------------------------------------+
  MAXIMUM EXTREME SCORE:     12 Persons

     SUMMARY OF 60 MEASURED (EXTREME AND NON-EXTREME) Persons
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       50.8      57.0        3.46     .80                               |
| S.D.        5.7        .0        1.74     .56                               |
| MAX.       57.0      57.0        6.23    1.86                               |
| MIN.       30.0      57.0         .07     .31                               |
|-----------------------------------------------------------------------------|
| REAL RMSE    .99  ADJ.SD    1.44  SEPARATION  1.45  Person RELIABILITY  .68 |
|MODEL RMSE    .98  ADJ.SD    1.44  SEPARATION  1.47  Person RELIABILITY  .68 |
| S.E. OF Person MEAN = .23                                                   |
+-----------------------------------------------------------------------------+
Person RAW SCORE-TO-MEASURE CORRELATION = .91
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .87

     SUMMARY OF 57 MEASURED (NON-EXTREME) Items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       41.5      48.0         .00     .62      1.00     .1   1.25     .3 |
| S.D.        6.8        .0        1.33     .24       .11     .4   1.60     .9 |
| MAX.       47.0      48.0        3.51    1.03      1.27    2.0   9.90    3.8 |
| MIN.       17.0      48.0       -1.67     .33       .75    -.9    .14   -1.0 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .68  ADJ.SD    1.14  SEPARATION  1.67  Item   RELIABILITY  .74 |
|MODEL RMSE    .66  ADJ.SD    1.15  SEPARATION  1.73  Item   RELIABILITY  .75 |
| S.E. OF Item MEAN = .18                                                     |
+-----------------------------------------------------------------------------+
  MINIMUM EXTREME SCORE:      6 Items
UMEAN=.000 USCALE=1.000

     SUMMARY OF 63 MEASURED (EXTREME AND NON-EXTREME) Items
+-----------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       42.1      48.0        -.28     .74                               |
| S.D.        6.7        .0        1.52     .43                               |
| MAX.       48.0      48.0        3.51    1.85                               |
| MIN.       17.0      48.0       -2.91     .33                               |
|-----------------------------------------------------------------------------|
| REAL RMSE    .86  ADJ.SD    1.26  SEPARATION  1.45  Item   RELIABILITY  .68 |
|MODEL RMSE    .85  ADJ.SD    1.26  SEPARATION  1.49  Item   RELIABILITY  .69 |
| S.E. OF Item MEAN = .19                                                     |
+-----------------------------------------------------------------------------+
```

Fig. 2. Summary statistic

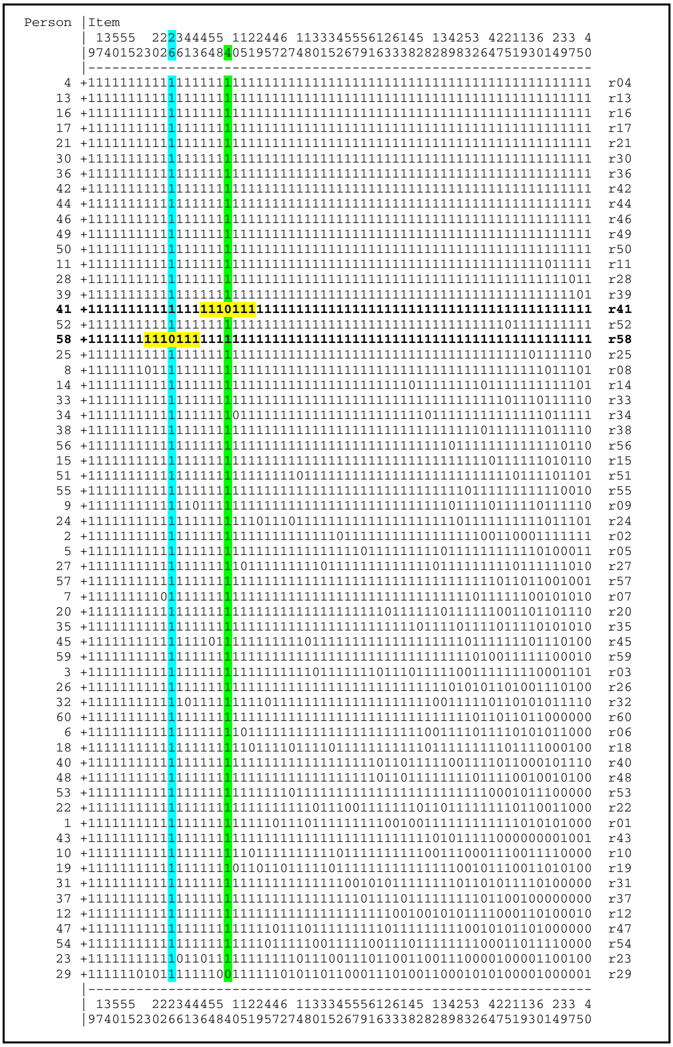
The Item Measure in Fig. 3 lists the family and criteria in an ascending order of their acceptability level. Close study reveals that the 6 items with extreme score or have minimum estimated *logit* measure are: *Availability of blog*; *Easy to read info*; *Clear layout of info; Family of Readability; Family of Info Representation; Family of Currency;* and *Appropriate explanatory text*.



```
+-----------------------------------------------------------------------------------------------+
|ENTRY    RAW                          MODEL|   INFIT    |  OUTFIT    |PTMEA|EXACT  MATCH|                         |
|NUMBER  SCORE   COUNT  MEASURE   S.E. |MNSQ   ZSTD|MNSQ   ZSTD|CORR.|  OBS%   EXP%|  Item                    |
|------------------------------------+----------+----------+-----+-----------+------------------------|
|     40      17       48      3.51       .34| .95    -.2|1.40    1.6|  .66| 75.0   74.4| 40-Must-have sound       |
|      5      23       48      2.84       .33| .93    -.5| .94    -.3|  .62| 68.8   70.4| 5-Relevant info          |
|     37      27       48      2.41       .33| .99     .0|1.10     .5|  .54| 66.7   69.9| 37-Info in diff format   |
|     24      28       48      2.30       .33| .90    -.7| .78  -1.0|  .59| 68.8   69.9| 24-Easy to remember add  |
|     39      28       48      2.30       .33| .91    -.7| .81    -.8|  .58| 72.9   69.9| 39-Must-have photos      |
|      1      29       48      2.19       .33|1.27    2.0|1.48    1.8|  .39| 62.5   70.0| 1-F.Accuracy             |
|     11      34       48      1.62       .35| .89    -.7| .76    -.7|  .51| 77.1   73.8| 11-Avail. Blog owner info|
|     19      34       48      1.62       .35| .86    -.9| .68  -1.0|  .53| 72.9   73.8| 19-Real-occurences info  |
|     33      34       48      1.62       .35|1.03     .3|1.01     .1|  .44| 68.8   73.8| 33-Technorati rank       |
|     60      34       48      1.62       .35|1.17    1.1|1.08     .3|  .39| 64.6   73.8| 60-Chat box              |
|     25      36       48      1.37       .36| .97    -.1|1.54    1.4|  .41| 79.2   77.1| 25-Emotional support     |
|     27      37       48      1.23       .37|1.15     .8| .91    -.1|  .37| 72.9   78.9| 27-Personal feel         |
|      6      38       48      1.08       .39|1.07     .4|1.18     .5|  .35| 79.2   80.8| 6-Originality            |
|     44      38       48      1.08       .39|1.01     .1| .97     .1|  .38| 83.3   80.8| 44-Interactivity         |
|     32      39       48       .93       .40| .91    -.4| .69    -.6|  .43| 85.4   82.6| 32-Rewarding experience  |
|     28      40       48       .76       .42|1.00     .1| .82    -.2|  .37| 83.3   84.3| 28-Surprises             |
|     53      40       48       .76       .42|1.06     .3| .93     .0|  .34| 83.3   84.3| 53-Readable font         |
|      8      42       48       .38       .46|1.03     .2|1.10     .4|  .30| 85.4   87.9| 8-Amount of info         |
|     12      42       48       .38       .46|1.04     .2| .79    -.1|  .31| 85.4   87.9| 12-Easy to understand    |
|     38      42       48       .38       .46|1.02     .2| .86     .0|  .31| 89.6   87.9| 38-Multimedia            |
|     49      42       48       .38       .46|1.08     .4|1.37     .7|  .25| 89.6   87.9| 49-Intuitive interface   |
|     13      43       48       .15       .50| .97     .0| .59    -.4|  .33| 87.5   89.7| 13-Informative           |
|     48      43       48       .15       .50| .98     .1|1.21     .5|  .28| 91.7   89.7| 48-Good use of colours   |
|     52      43       48       .15       .50| .93    -.1| .66    -.3|  .33| 91.7   89.7| 52-Legibility            |
|     16      44       48      -.13       .55|1.13     .4| .78     .0|  .23| 91.7   91.7| 16-Link to info          |
|     23      44       48      -.13       .55|1.10     .4| .77     .0|  .25| 91.7   91.7| 23- Cognitive advancement|
|     63      44       48      -.13       .55| .75    -.5| .37    -.7|  .38| 91.7   91.7| 63-Trackback             |
|      7      45       48      -.46       .62|1.18     .5|1.30     .6|  .14| 93.8   93.8| 7- F.Completeness        |
|     14      45       48      -.46       .62| .81    -.3|1.10     .4|  .28| 93.8   93.8| 14-Objective info        |
|     18      45       48      -.46       .62|1.11     .4| .85     .2|  .20| 93.8   93.8| 18-Real time info        |
|     30      45       48      -.46       .62|1.10     .4| .74     .1|  .21| 93.8   93.8| 30-Reputation of blog    |
|     31      45       48      -.46       .62| .90     .0| .79     .1|  .26| 93.8   93.8| 31- Reputation of blogger|
|     35      45       48      -.46       .62| .97     .1| .54    -.2|  .27| 93.8   93.8| 35-Exciting content      |
|     42      45       48      -.46       .62|1.00     .2| .84     .2|  .24| 93.8   93.8| 42-Ease of ordering      |
|     56      45       48      -.46       .62| .87    -.1| .48    -.3|  .30| 93.8   93.8| 56-Blog responsiveness   |
|     57      45       48      -.46       .62| .87    -.1| .48    -.3|  .30| 93.8   93.8| 57-Ease of info access   |
|     59      45       48      -.46       .62| .91     .0| .82     .2|  .26| 93.8   93.8| 59-Blogroll              |
|     61      45       48      -.46       .62| .93     .0| .43    -.4|  .30| 93.8   93.8| 61-Comment field         |
|      4      46       48      -.92       .75| .91     .1|7.69    2.8|  .10| 95.8   95.9| 4-Reliable source        |
|     10      46       48      -.92       .75|1.13     .4|2.48    1.3|  .08| 95.8   95.9| 10-Appropriate level     |
|     15      46       48      -.92       .75|1.14     .4|1.34     .7|  .11| 95.8   95.9| 15-Provide info sources  |
|     21      46       48      -.92       .75|1.08     .3| .72     .1|  .17| 95.8   95.9| 21-F.Engaging            |
|     29      46       48      -.92       .75|1.12     .4|1.44     .7|  .12| 95.8   95.9| 29-F.Reputation          |
|     45      46       48      -.92       .75|1.08     .3| .76     .2|  .17| 95.8   95.9| 45-F.Visual design       |
|     47      46       48      -.92       .75|1.05     .3| .57     .0|  .20| 95.8   95.9| 47-Clear layout          |
|     62      46       48      -.92       .75| .83    -.1| .36    -.3|  .28| 95.8   95.9| 62-Search tool           |
|      2      47       48     -1.67      1.03| .80     .1| .14    -.6|  .25| 97.9   97.9| 2-Correct info           |
|      3      47       48     -1.67      1.03|1.11     .4|4.61    1.9| -.02| 97.9   97.9| 3-Reliable info          |
|     20      47       48     -1.67      1.03| .80     .1| .14    -.6|  .25| 97.9   97.9| 20-Up-to-date            |
|     22      47       48     -1.67      1.03|1.09     .4|1.52     .8|  .06| 97.9   97.9| 22- Appreciate comments  |
|     26      47       48     -1.67      1.03|1.12     .4|9.90    3.8| -.11| 97.9   97.9| 26-Fun                   |
|     36      47       48     -1.67      1.03| .91     .2| .21    -.4|  .22| 97.9   97.9| 36-Fresh perspective     |
|     41      47       48     -1.67      1.03|1.08     .4|1.04     .5|  .09| 97.9   97.9| 41-F.Navigation          |
|     43      47       48     -1.67      1.03|1.10     .4|2.41    1.2|  .03| 97.9   97.9| 43-Easy to navigate      |
|     46      47       48     -1.67      1.03| .91     .2| .21    -.4|  .22| 97.9   97.9| 46-Attractive layout     |
|     54      47       48     -1.67      1.03|1.09     .4|1.52     .8|  .06| 97.9   97.9| 54-F.Blog accessibility  |
|     58      47       48     -1.67      1.03| .80     .1| .14    -.6|  .25| 97.9   97.9| 58-F.Blog Tech Features  |
|      9      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 9-Appropriate exp. text  |
|     17      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 17-F.Currency            |
|     34      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 34-F.Info representation |
|     50      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 50-F.Readability         |
|     51      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 51-Easy to read info     |
|     55      48       48     -2.91      1.85|  MINIMUM ESTIMATED MEASURE  |             | 55-Availability of blog  |
|------------------------------------+----------+----------+-----+-----------+------------------------|
| MEAN    42.1    48.0      -.28       .74|1.00     .1|1.25     .3|     | 88.3   88.8|                          |
| S.D.     6.7      .0      1.52       .43| .11     .4|1.60     .9|     | 10.2    9.3|                          |
+-----------------------------------------------------------------------------------------------+
```

Fig. 3. Item Measure

By looking at the Point-Measure Correlation (see column titled PTMEA CORR.), it can be found that 15 items are in the acceptable range; $0.32<x<0.80$. On the other hand, 37 items fall outside the range. A further verification for these items is done by looking at the OUTFIT column for MNSQ, y value; $0.5<y<1.5$. Sixteen items are found beyond this parameter. Further check on the Z-Std value, $-2<z<2$; shows *Reliable source* and *Fun* are beyond the upper limit, +2. Counter check against the Guttman scalogram (see Fig. 4) indicates that the two items, *Reliable source* (item 4) and *Fun* (item 26) have been under rated by respondent 41 and respondent 58 respectively.



```
Person |Item
       | 13555  222344455 1122446 1133345556126145 134253 4221136 233 4
       |974015230266136484051957274801526791633382828983264751930149750
       |---------------------------------------------------------------
     4 +111111111111111111111111111111111111111111111111111111111111111  r04
    13 +111111111111111111111111111111111111111111111111111111111111111  r13
    16 +111111111111111111111111111111111111111111111111111111111111111  r16
    17 +111111111111111111111111111111111111111111111111111111111111111  r17
    21 +111111111111111111111111111111111111111111111111111111111111111  r21
    30 +111111111111111111111111111111111111111111111111111111111111111  r30
    36 +111111111111111111111111111111111111111111111111111111111111111  r36
    42 +111111111111111111111111111111111111111111111111111111111111111  r42
    44 +111111111111111111111111111111111111111111111111111111111111111  r44
    46 +111111111111111111111111111111111111111111111111111111111111111  r46
    49 +111111111111111111111111111111111111111111111111111111111111111  r49
    50 +111111111111111111111111111111111111111111111111111111111111111  r50
    11 +111111111111111111111111111111111111111111111111111111111011111  r11
    28 +111111111111111111111111111111111111111111111111111111111111011  r28
    39 +111111111111111111111111111111111111111111111111111111111111101  r39
    41 +111111111111111110111111111111111111111111111111111111111111111  r41
    52 +111111111111111111111111111111111111111111111111111101111111111  r52
    58 +111111111101111111111111111111111111111111111111111111111111111  r58
    25 +111111111111111111111111111111111111111111111111111110111111110  r25
     8 +111111101111111111111111111111111111111111111111111111111011101  r08
    14 +111111111111111111111111111111111111111101111111101111111111101  r14
    33 +111111111111111111111111111111111111111111111111111101110111110  r33
    34 +111111111111111111011111111111111111111110111111111111110111111  r34
    38 +111111111111111111111111111111111111111111111111101111110111110  r38
    56 +111111111111111111111111111111111111111111111011111111111111010  r56
    15 +111111111111111111111111111111111111111111111111110111111010110  r15
    51 +111111111111111111111111110111111111111111111111111110111110101  r51
    55 +111111111111111111111111111111111111111111111110111111111110010  r55
     9 +111111111111101111111111111111111111111111111011110111110111110  r09
    24 +111111111111111111111011101111111111111111111101111111111011101  r24
     2 +111111111111111111111111111111101111111111111111100110001111111  r02
     5 +111111111111111111111111111111111101111111110111111111110100011  r05
    27 +111111111111111111101111111110111111111111101111111110111111010  r27
    57 +111111111111111111111111111111111111111111111111111011011001001  r57
     7 +111111111011111111111111111111111111111111111111011111100101010  r07
    20 +111111111111111111111111111111111111101111110111111001101101110  r20
    35 +111111111111111111111111111111111111111110111101111011110101010  r35
    45 +111111111111111011111111111011111111111111111110111111101110100  r45
    59 +111111111111111111111111111111111111111111111111010011111100010  r59
     3 +111111111111111111111111111101111111011101111100111111110001101  r03
    26 +111111111111111111111111111111111111111111111010101101001110100  r26
    32 +111111111111011111111101111111111111111111110011111011010101110  r32
    60 +111111111111111111111111111111111111111111111111011011011000000  r60
     6 +111111111111111111101111111111111111111111001111011110101011000  r06
    18 +111111111111111111110111101111011111111111011111111110111000100  r18
    40 +111111111111111111111111111111111111011011111001111011000101110  r40
    48 +111111111111111111111111111111111111011011111111011110010010100  r48
    53 +111111111111111111111111101111111111111111111111110001011100000  r53
    22 +111111111111111111111111111101110011111110110111111110110011000  r22
     1 +111111111111111111111110111011111111100100111111111111010101000  r01
    43 +111111111111111111111111111111111111111111101011111000000001001  r43
    10 +111111111111111111110111111111101111111111001110001110011110000  r10
    19 +111111111111111111011011011110111111111111111100101110011010100  r19
    31 +111111111111111111111111111111110010101111111101101011110100000  r31
    37 +111111111111111111111111111111111101111011111111101100100000000  r37
    12 +111111111111111111111111111111111111110010010101111000110100010  r12
    47 +111111111111111111111110111011111111101111111110010101101000000  r47
    54 +111111111111111111111101111100111110011011111111100011011110000  r54
    23 +111111111110110111111111110111001110110011001110000100001100100  r23
    29 +111111010111111100111111010110110001110100110001010100001000001  r29
       |---------------------------------------------------------------
       | 13555  222344455 1122446 1133345556126145 134253 4221136 233 4
       |974015230266136484051957274801526791633382828983264751930149750
```

Fig. 4. Guttman scalogram

One possible reason is that they could have been careless in attempting their decisions which lead to such a grossly under rated work. After verifying that the Infit value (see INFIT column in Fig. 3) is within range (MNSQ: $0.5<y<1.5$; Z-Std: $-2<z<2$), the two misfits are acceptable.

As stated in the objective, there are two different aspects to this analysis, firstly the definition of family and secondly the assigning of criteria to the respective families. In order to analyze experts' views and comments from the open question provided in the content validity test, the percentage of the probability of the two aforementioned aspects to be agreed is calculated based on the *logit* measure. This is to decide whether to review them or not. A threshold value to 70% is set in line with the standard threshold limit of Cronbach Alpha. It can then be construed as follows:

- Definition of family and the assigning of criteria with percentage of probability to be agreed of more than 70% will be accepted without being reviewed.
- Definition of family and the assigning of criteria with percentage of probability to be agreed of less than 70% will be reviewed if comments are provided by the experts. The family will then be redefined whereas the criteria will be discarded or amended, if required.

The results for the 11 families are presented in Table 1. The summary shows that the definitions of 9 families are agreeable by the experts with the percentage of probability to be agreed is between 70% to 95%. On the other hand, the definition of *Family of Accuracy* and *Family of Completeness* need to be reviewed, with the percentage of probability to be agreed below 70%. However, the definition of the *Family of Completeness* is accepted without being reviewed because there is no comment available for it. On the other hand, *Family of Accuracy* has been redefined as suggested by the experts. See Table 3 for the accepted definitions of the 11 families.

Table 1. Percentage of possibility to be agreed for the definitions of 11 families

Family		P(Θ) (%)	Family		P(Θ) (%)
1	Accuracy	10.07	7	Blog Accessibility	84.16
2	Completeness	61.30	8	Blog Technical Features	84.16
3	Engaging	71.50	9	Currency	94.83
4	Reputation	71.50	10	Info Representation	94.83
5	Visual Design	71.50	11	Readability	94.83
6	Navigation	84.16			

The findings for the assigning of criteria to the respective families are shown in Table 2. It can be seen that 16 criteria (percentage of possibility to be agreed > 70%) remain in their respective families. Based on the findings, there are 36 criteria that require to be reviewed. However, there is no comment provided for the 31 criteria, means that they remain in their respective families. Yet, there are 5 criteria have been revisited; (1) *Relevant info* from *Family of Accuracy*, (2) *Easy to remember address* from *Family of Engaging*, (3) *Must-have sounds*, (4) *Info displayed in different format*, and (5) *Must have photos*. The later 3 criteria are from *Family of Info Representation*. Consequently, as suggested by the experts, the following actions have been taken for each of them:

- *Relevant info* is deleted from its family and has been transferred to the *Family of Completeness.*
- *Easy to remember address* is replaced by *Memorable content.*
- *Info displayed in different format* is eliminated from its family for sharing the same meaning as *Multimedia.*
- *Must-have photos* is discarded from its family for it is an integral part of *Multimedia.*
- *Must-have sounds* is removed from its family for it is also an integral part of *Multimedia.*

Table 2. Percentage of probability to be agreed for the assigning of 52 criteria

Family		P(Θ) (%)	Family		P(Θ) (%)
1	Must-have sound	2.90	27	Objective info	61.30
2	Relevant info	5.52	28	Real time info	61.30
3	Info in different format	8.24	29	Reputation of blog	61.30
4	Easy to remember address	9.11	30	Reputation of blogger	61.30
5	Must-have photos	9.11	31	Exciting content	61.30
6	Availability of blog owner info	16.52	32	Ease of ordering	61.30
7	Real-occurrence info	16.52	33	Blog responsiveness	61.30
8	Technorati rank	16.52	34	Ease of information access	61.30
9	Chat box	16.52	35	Blogroll	61.30
10	Emotional support	20.26	36	Comment field	61.30
11	Personal feel	22.62	37	Reliable source	71.50
12	Originality	25.35	38	Appropriate level of content	71.50
13	Interactivity	25.35	39	Provide information source	71.50
14	Rewarding experience	28.29	40	Clear layout of info	71.50
15	Surprises	31.86	41	Search tool	71.50
16	Readable font	31.86	42	Correct info	84.16
17	Amount of info	40.61	43	Reliable info	84.16
18	Easy to understand	40.61	44	Up-to-date	84.16
19	Multimedia	40.61	45	Appreciate comments	84.16
20	Intuitive interface	40.61	46	Fun	84.16
21	Informative	46.26	47	Fresh perspective	84.16
22	Good use of colours	46.26	48	Easy to navigate	84.16
23	Legibility	46.26	49	Attractive layout	84.16
24	Link to info	53.25	50	Appropriate explanatory text	94.83
25	Cognitive advancement	53.25	51	Easy to read info	94.83
26	Trackback	53.25	52	Availability of blog	94.83

See Table 3 for the final assigning of the 49 criteria to the 11 families concerned. They will then be used in the construction of questionnaire for measuring the acceptability of criteria for blog quality.

Table 3. Final result of content validity test

Family		Definition	Quality criteria	
1	Accuracy	The extent to which information is exact and correct, certified as being free-of-error.	1	Correct information
			2	Reliable info
			3	Reliable source
			4	Originality
2	Completeness/ Comprehensiveness of Info	The extent to which the information provided is sufficient.	5	Amount of information
			6	Appropriate explanatory text
			7	Appropriate level of content
			8	Availability of blog owner information
			9	Easy to understand information
			10	Informative
			11	Links to information
			12	Objective information
			13	Providing information sources
			14	Relevant info
3	Currency, Timeliness, Update	The extent to which the blog provides non-obsolete information.	15	Real time info
			16	Real-occurrence info
			17	Up-to-date info
4	Engaging	The extent to which the blog can attract and retain readers.	18	Appreciation for readers' comments
			19	Cognitive advancement
			20	Emotional support
			21	Fun
			22	Surprises
			23	Personal feel
			24	Memorable content
5	Reputation	The extent to which the information is trusted or highly regarded in terms of their source or content.	25	Reputation of blog
			26	Reputation of bloggers
			27	Rewarding experiences
			28	Technorati rank
6	Info Representation	The way information is presented, maybe in different formats/ media with customized displays.	29	Exciting content
			30	Fresh perspective
			31	Multimedia
7	Navigation	The extent to which readers can move around the blog and retrieve information easily.	32	Ease of ordering
			33	Easy to navigate
			34	Interactivity

Table 3. (*continued*)

Family		Definition	Quality criteria	
8	Visual Design	Visual appearances that can attract readers.	35	Attractive layout
			36	Clear layout of info
			37	Good use of colours
			38	Intuitive interface
9	Readability	Ability to comprehend the meaning of words or symbols.	39	Easy to read info
			40	Legibility
			41	Readable font/ text
10	Blog Accessibility	The extent to which the blog can be accessed faster and easier.	42	Availability of info
			43	Blog responsiveness
			44	Ease of information access
11	Blog Technical Features	Features such as search tools, chat box, blogroll, and comment field.	45	Blogroll
			46	Chat box
			47	Comment field
			48	Search tool
			49	Trackback

5 Conclusion and Future Work

This paper has described the content validity test to confirm whether the prescribed family definitions and blog quality criteria assigned are agreeable to the reviewers by means of online survey.

In conclusion, this study confirms the content validity of the 49 criteria in the 11 families for blog quality through a reliable and valid content validity test. These criteria will then be used as a basis in the construction of the instrument to measure the acceptability of blog quality criteria based on users' perception. It is also found that, despite the small sample size, Rasch Measurement Model is an effective tool in assessing content validity accurately and fast.

The content validity test is crucial to ensure that the content of our questionnaire is significant for meaningful measurement. Means that the instrument's content is essential to identify the important criteria from the blog readers' viewpoints in determining a blog quality. Consequently, it is an initial step towards the achieving a valid blog quality model. The model can be used as guidelines for blog readers to verify whether the visited blog is of quality or not. Besides, the model can help bloggers to promote readers' satisfaction.

However, this study does not establish the construct validity and criterion validity of the measurement instrument. Also, there is no evidence as to whether it is reliable or not. Therefore, as for future work, we plan to continue identifying the three aforementioned aspects in an effort to develop an accurate measurement instrument for getting a precise and correct blog quality model.

References

1. Lenhart, A., Fox, S.: A portrait of the internet's new storytellers (2006)
2. Vanier, F.: World Broadband Statistics: Q4 2009 (2010)
3. Tan, J.-E., Ibrahim, Z.: Blogging and Democratization in Malaysia. A New Civil Society in the Making. SIRD, Petaling Jaya (2008)
4. Kitchenham, B.A., Pfleeger, S.L.: Personal Opinion Surveys. In: Shull, F., Singer, J., Sjøberg, D.I.K. (eds.) Guide to Advanced Empirical Software Engineering, pp. 71–92. Springer, London (2008)
5. Abdul Aziz, A., Mohamed, A., Arshad, N., Zakaria, S., Zaharim, A., Ahmad Ghulman, H., Masodi, M.S.: Application of Rasch Model in validating the construct of measurement instrument. International Journal of Education and Information Technologies 2(2) (2008)
6. Rasch, G.: Weblogs models for some intelligence and Student test. The University of Chicago Press, Chicago (1980)
7. Masodi, M.S., Abdul Aziz, A., Mohamed, A., Arshad, N., Zakaria, S., Ahamd Ghulman, H.: Development of Rasch-based Descriptive Scale in profiling Information Professionals' Competency. In: IEEE IT Simposium, Kuala Lumpur, pp. 329–333 (2008)
8. Sick, J.: Rasch Measurement in Language Education Part 3: The family of Rasch Models. Shiken. JALT Testing & Evaluation SIG Newsletter 13(1), 4–10 (2009)
9. Wright, B.D.: Rasch Model from Counting Right Answers: Raw Scores as Sufficient Statistics. Rasch Measurement Transactions 3(2), 62 (1989)
10. Wright, B.D., Mok, M.M.C.: An overview of the family of Rasch measurement models. In: Everett, J., Smith, V., Smith, R.M. (eds.) Introduction to Rasch Measurement: Theory, Models, and Applications, p. 979 (2004)
11. Abdul Aziz, A.: Rasch Model Fundamentals: Scale Construct and Measurement Structure. Perpustakaan Negara Malaysia, Kuala Lumpur (2010)
12. Bond, T.V., Fox, C.M.: Applying The Rasch Model: Fundamental Measurement in the Human Sciences, 2nd edn. Lawrence Erlbaum Associates, New Jersey (2007)

Super Attribute Representative for Decision Attribute Selection

Rabiei Mamat[1], Tutut Herawan[2], and Mustafa Mat Deris[3]

[1] Department of Computer Science, Universiti Malaysia Terengganu
Gong Badak 21030 Kuala Terengganu, Terengganu, Malaysia
rab@umt.edu.my
[2] Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang
Lebuhraya Tun Razak, Gambang 26300, Kuantan Pahang, Malaysia
tutut@ump.edu.my
[3] Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat 86400, Johor, Malaysia
mmustafa@uthm.edu.my

Abstract. Soft set theory proposed by Molodstov is a general mathematic tool for dealing with uncertainties. Recently, several algorithms had been proposed for decision making using soft set theory. However, these algorithms still concern on a Boolean-valued information system. In this paper, Support Attribute Representative (SAR), a soft set based technique for decision making in categorical-valued information system is proposed. The proposed technique has been tested on two datasets. The results of this research will provide useful information for decision makers to handle categorical datasets.

Keywords: Information system; Data mining; Soft set theory; Decision attributes selections.

1 Introduction

In today's fast moving world, decision making is a very critical issues. Good decision making is aiding by the good information. Unfortunately, some information is uncertain. Handling uncertain data is very important because in reality, there are lots real life problems in which still involve uncertain data, for example in field of engineering, medical, social, medical sciences and etc [1]. There are several theories, such as probabilities theories, theory of fuzzy set, theory of rough set and etc, which can be considered as the mathematical tools for dealing with uncertainties.

The theory of soft set proposed by Molodtsov [2] 1999 as a new way for managing uncertain data. Molodtsov pointed out that one of the main advantages of soft set theory is that it is free from the inadequacy of the parameterization tools, unlike in the theories mentioned above. The soft set theory uses parameterization sets, as its main vehicles for problem solving, which makes it very convenient and easy to apply in practice. Therefore, many applications based on soft set theory have already been demonstrated by Molodtsov [2], such as the smoothness of functions, game theory, operations research, Riemann integration, Perron integration, probability theory, and

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 137–147, 2011.

measurement theory. Presently, great progresses of study on soft set theory have been made. Maji *et al.* [3] firstly introduced some definitions of the related operations on soft sets. Ali *et al.* [4] took into account some errors of former studies and put forward some new operations on soft sets. As for practical applications of soft set theory, great progress has been achieved. Maji *et al.* [1] employed soft sets to solve the decision-making problem. Roy and Maji [5] presented a novel method of object recognition from an imprecise multi-observer data to deal with decision making based on fuzzy soft sets, which was revised by Kong *et al.* [6]. Feng *et al.* [7] showed an adjustable approach to fuzzy soft set based decision making by means of level soft sets. It is worthwhile to mention that some effort has been done to such issues concerning reduction of soft sets. Chen *et al.* [8] pointed out that the conclusion of soft set reduction offered in [1] was incorrect, and then present a new notion of parameterization reduction in soft sets in comparison with the definition to the related concept of attributes reduction in rough set theory. The concept of normal parameter reduction is introduced in [9], which overcome the problem of suboptimal choice and added parameter set of soft sets. An algorithm for normal parameter reduction is also presented in [9]. However, these soft set-based algorithms still concern on a Boolean-valued information system.

In this paper, Support Attribute Representative (SAR), a soft set based technique for decision making in categorical-valued information system is proposed. The proposed technique has been tested on two datasets. The results of this research will provide useful information for decision makers to handle categorical datasets.

The rest of the paper is organized as follows. Section 2 described the soft set theory. In section 3, the proposed technique is describe. An experiment and analysis is described in section 4. Finally, conclusion of this work described in section 5.

2 Soft Set Theory

An *information system* as in is a 4-tuple (quadruple) $S=(U,A,V,f)$, where $U=\{u_1,u_2,u_3,\cdots,u_{|U|}\}$ is a non-empty finite set of objects, $A=\{a_1,a_2,a_3,\cdots,a_{|A|}\}$ is a non-empty finite set of attributes, $V=\bigcup_{a\in A}V_a$, V_a is the domain (value set) of attribute *a*, $f:U\times A\to V$ is an information function such that $f(u,a)\in V_a$, for every $(u,a)\in U\times A$, called information (knowledge) function. An information system can be intuitively expressed in terms of an information table (see Table 1).

Table 1. An information system

U	a_1	$\cdots$	a_k	$\cdots$	$a_{\vert A\vert}$
u_1	$f(u_1,a_1)$	$\cdots$	$f(u_1,a_k)$	$\cdots$	$f(u_1,a_{\vert A\vert})$
u_2	$f(u_2,a_1)$	$\cdots$	$f(u_2,a_k)$	$\cdots$	$f(u_2,a_{\vert A\vert})$
u_3	$f(u_3,a_1)$	$\cdots$	$f(u_3,a_k)$	$\cdots$	$f(u_3,a_{\vert A\vert})$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$u_{\vert U\vert}$	$f(u_{\vert U\vert},a_1)$	$\cdots$	$f(u_{\vert U\vert},a_k)$	$\cdots$	$f(u_{\vert U\vert},a_{\vert A\vert})$

In many applications, there is an outcome of classification that is known. This a posteriori knowledge is expressed by one (or more) distinguished attribute called decision attribute; the process is known as supervised learning. An information system of this kind is called a decision system. A *decision system* is an information system of the form $D = (U, A \cup \{d\}, V, f)$, where $d \notin A$ is the decision attribute.

Throughout this section U refers to an initial universe, E is a set of parameters, $P(U)$ is the power set of U and $A \subseteq E$.

Definition 1. (See [2].) *A pair* (F, A) *is called a soft set over* U *, where* F *is a mapping given by*

$$F : A \to P(U).$$

In other words, a soft set over U is a parameterized family of subsets of the universe U. For $\varepsilon \in A$, $F(\varepsilon)$ may be considered as the set of ε-elements of the soft set (F, A) or as the set of ε-approximate elements of the soft set. Clearly, a soft set is not a (crisp) set.

As an illustration, let consider a soft set which describes the 'attractiveness of programming language' that Mr. X is considering to use. Let assume that there are ten programming languages in the universe U that are under consideration by Mr. X such that $U = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}\}$, and is a set of decision parameter where $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$. Every $e \in E$ representing an element of attractiveness such that e_1 representing an element of 'system', e_2 representing an element of 'education', e_3 representing an element of 'web', e_4 representing an element of 'procedural', e_5 representing an element of 'object oriented', e_6 representing an element of 'reflective' and e_7 representing an element of 'functional'. Consider the mapping $F : E \to P(U)$ given by "programming language (.)", where (.) is to be filled in by one of parameters $e \in E$. Suppose that $F(e_1) = \{p_1, p_3, p_4, p_{10}\}$, $F(e_2) = \{p_2, p_7, p_8\}$, $F(e_3) = \{p_5, p_8, p_9\}$, $F(e_4) = \{p_1, p_2, p_3, p_4, p_7\}$, $F(e_5) = \{p_1, p_4, p_5, p_6, p_7, p_8, p_9\}$, $F(e_6) = \{p_5, p_8, p_9\}$ and $F(e_7) = \{p_5, p_8, p_9\}$. As for example, $F(e_1)$ means system programming languages, whose functional value is the set $\{p_1, p_3, p_4, p_{10}\}$. Thus we can view the soft set (F, E) as a collection of approximations as illustrated below:

$$(F, E) = \left\{ \begin{array}{l} e_1 = \{p_1, p_3, p_4, p_{10}\} \\ e_2 = \{p_2, p_7, p_8\} \\ e_3 = \{p_5, p_8, p_9\} \\ e_4 = \{p_1, p_2, p_3, p_4, p_7\} \\ e_5 = \{p_1, p_4, p_5, p_6, p_7, p_8, p_9\} \\ e_6 = \{p_5, p_8, p_9\} \\ e_7 = \{p_5, p_8, p_9\} \end{array} \right\}$$

Proposition 1. *If (F,E) is a soft set over the universe U, then (F,E) is a Boolean-valued information system* $S = (U, A, V_{\{0,1\}}, f)$

Proof. Let (F,E) be a soft set over the universe U, we define a mapping $F=\{f_1,f_2,..,f_n\}$ where

$$f_i : U \to V_i \qquad and \qquad f_i(x)=\begin{cases}1, & x\in F(e_i)\\ 0, & x\notin F(e_i)\end{cases}, for \quad 1\le x\le |A|$$

Hence, if $A=E,\ V=\bigcup_{e_i\in A} V_{e_i}$ where $V_{e_i}=\{0,1\}$ then the soft set (F,E) can be considered as a Boolean-valued information system $S=(U,A,V_{\{0,1\}},f)$. Therefore, a soft set (F,E) can be represented in the form of binary table. As can be seen in the Table 2, '1' denote the presence of the described parameters, while '0' mean the parameter is not part of the description of the programming languages attractiveness.

Table 2. Tabular representation of a soft set (F,E)

(F,E)	e_1	e_2	e_3	e_4	e_5	e_6	e_7
p_1	1	0	0	1	1	0	0
p_2	0	1	0	1	0	0	0
p_3	1	0	0	1	0	0	0
p_4	1	0	0	1	1	0	0
p_5	0	0	1	0	1	1	1
p_6	0	0	0	0	1	0	0
p_7	0	1	0	1	1	0	0
p_8	0	1	1	0	1	1	1
p_9	0	0	1	0	1	1	1
p_{10}	1	0	0	0	1	0	0

For multi-valued information system, it needs to convert into *multi-soft sets* [10]. It is based on the notion of a decomposition of a multi-valued information system. Let $S=(U,A,V,f)$ be a multi-valued information system and $S^i=(U,a_i,V_{ai},f)$, $i=1,2,\cdots,|A|$ be the $|A|$ binary-valued information systems. From Proposition 1, we have

$$S=(U,A,V,f)=\begin{cases} S^1=(U,a_1,V_{\{0,1\}},f) & \Leftrightarrow (F,a_1)\\ S^2=(U,a_2,V_{\{0,1\}},f) & \Leftrightarrow (F,a_2)\\ \quad\vdots & \quad\vdots \quad \vdots\\ S^{|A|}=(U,a_{|A|},V_{\{0,1\}},f) & \Leftrightarrow (F,a_{|A|}) \end{cases}$$
$$=\left((F,a_1),(F,a_2),\cdots,(F,a_{|A|})\right)$$

We define $(F,E)=\left((F,a_1),(F,a_2),\cdots,(F,a_{|A|})\right)$ as a multi-soft set over universe U representing a multi-valued information system $S=(U,A,V,f)$.

Example 1. We consider the following data to illustrate the concept of multi-soft sets.

Table 3. Multi-value information system

Object	Volume	Material	Location
1	High	Hard	Pacific
2	High	Hard	Midwest
3	High	Medium	East Coast
4	High	Soft	Northern
5	Low	Soft	Pacific
6	Low	Medium	Midwest

The Dataset consist of three categorical-valued attributes so-called Volume, Material and Location. Attribute Volume only has two categories i.e. High and low. Attribute Material has 3 categories i.e. Hard, Medium and Soft. Meanwhile, attribute Location contains four categories i.e. Pacific, Midwest, East-Coast and Northern. The multi-soft sets representing the above information system is as follow.

$$(F,E)=\left\{\begin{array}{l}(F,Volume_{High})=\{1,2,3,4\}\\(F,Volume_{Low})=\{5,6\}\\(F,Material_{Hard})=\{1,2\}\\(F,Material_{Medium})=\{3,6\}\\(F,Material_{Soft})=\{4,5\}\\(F,Location_{Pacific})=\{1,5\}\\(F,Location_{Midwest})=\{2,6\}\\(F,Location_{East-coast})=\{3\}\\(F,Location_{Northern})=\{4\}\end{array}\right\}$$

3 Super Attribute Representative (SAR)

Throughout this section a pair (F,A) refers to multi-soft sets of universe U representing a categorical-valued information system $S=(U,A,V,f)$.

Definition 2 (Support). *Let* (F,A) *is a multi-soft set over the universe* U*, where* $(F,a_i)..(F,a_{|A|})\subseteq(F,A)$, *and* $(F,a_{i_1})..(F,a_{|a_i|})\subseteq(F,a_i)$. *The support of* (F,a_{i_j}) *by* (F,a_{i_k}) *denoted by* $Sup_{(F,a_{i_k})}(F,a_{i_j})$ *is defined as*

$$Sup_{(F,a_{i_k})}(F,a_{i_j})=\frac{\left|(F,a_{i_j})\cap(F,a_{i_k})\right|}{\left|(F,a_{i_j})\cup(F,a_{i_k})\right|},\forall(F,a_{i_j})\in(F,A);1\le i\le|A|;1\le j,k\le|a_i|$$

Definition 3 (Total Support). *The summation of all support for* (F,a_{j_k}) *denoted by* $TSup(F,a_{j_k})$ *is defined as*

$$TSup\left(F, a_{j_k}\right) = \sum^{|A|} Sup_{\left(F, a_{m_n}\right)}\left(F, a_{j_k}\right), \qquad \forall\left(F, a_{j_k}\right) \in (F, A), 1 \le n \le |A|; 1 \le n \le |a_m|$$

Definition 4 (Attribute Representation). *The soft set in* (F, A) *that have maximum total support in their domain is known as attribute representative denoted by* $\mathrm{Rep}(A_i)$ *is defined as*

$$\mathrm{Rep}(A_i) = Max\left(TSup(\left(F, a_{i_1}\right),.., TSup\left(F, a_{i_{|A_i|}}\right)\right))$$

Definition 5 (Super Attribute Representative). *Attribute Representative with maximum value is known as Super Attribute Representative* (SAR), *defined as*

$$SAR = \begin{cases} Max\left(Mode\left(\mathrm{Re}\,p(A_i),.., \mathrm{Re}\,p\left(A_{|A|}\right)\right)\right), \textit{Multiple} \quad \textit{Occurences} \\ Max\left(\mathrm{Re}\,p(A_i),.., \mathrm{Re}\,p\left(A_{|A|}\right)\right), \quad \textit{No} \quad \textit{Multiple} \quad \textit{Occurence} \end{cases}$$

Next, the SAR algorithm is presented as follows.

```
Algorithm SAR
Input: Categorical-valued Dataset
Output: Decision Attribute
Begin
 1. Calculate Support and Total Support
      For i = all categories
         For j=all categories
             Intersection=Sum(And(Data(:,i),Data(:,j))
             Union = Sum(Or(Data(:,i),Data(:,j)
             Support i=Intersection / Union
             Total Support I += Support i
         End
      End

 2. Determine Representative of Domain $A_i$
    Size(R)=[row,column]
    For i=1:row
      If Mode([$A_i$])
         Rep($A_i$)=Large([$A_i$])
      else
       Rep($A_i$)=large([$A_i$])
      End

 3. Determine SAR of Domain A
     If mode(Rep())
      SAR=SAR(Rep())
     Else
        SAR = Large(Rep())
      end
```

Fig. 1. Algorithm SAR for Selecting Decision Attribute

3.1 Computational Complexity

Suppose that in an information system U , there are n objects, m attributes and l numbers of distinct values of each attribute. Process to determining an elementary set of all attributes needs nm . The computation of calculating the support of all subsets of U having different value of a_i is $n^2 l$. Meanwhile, by taking the worst-case, the process to determining the SAR is 1. Therefore, the computational complexity of SAR techniques is polynomial $O(n^2 l + nm + 1)$.

3.2 Example

In this section, an example of SAR implementation will be presented. Information system in Table 3 will be used as an input. All steps in this example are as the algorithm in Figure 1. Support for each soft set is computed using Definition 2. For example, all supports of $(F, Volume_{High})$ are calculated as follows.

a. Support by $(F, Volume_{High})$

$$Sup_{(F, Volume_{High})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{1,2,3,4\}|}{|\{1,2,3,4\} \cup \{1,2,3,4\}|} = \frac{4}{4} = 1$$

b. Support by $(F, Volume_{Low})$

$$Sup_{(F, Volume_{Low})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{5,6\}|}{|\{1,2,3,4\} \cup \{5,6\}|} = \frac{0}{6} = 0$$

c. Support by $(F, Material_{Hard})$

$$Sup_{(F, Material_{Hard})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{1,2\}|}{|\{1,2,3,4\} \cup \{1,2\}|} = \frac{2}{4} = 0.5$$

d. Support by $(F, Material_{Medium})$

$$Sup_{(F, Material_{Medium})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{3,6\}|}{|\{1,2,3,4\} \cup \{3,6\}|} = \frac{1}{5} = 0.2$$

e. Support by $(F, Material_{Soft})$

$$Sup_{(F, Material_{Soft})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{4,5\}|}{|\{1,2,3,4\} \cup \{4,5\}|} = \frac{1}{5} = 0.2$$

f. Support by $(F, Location_{Pacific})$

$$Sup_{(F, Location_{Pacific})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{1,5\}|}{|\{1,2,3,4\} \cup \{1,5\}|} = \frac{1}{5} = 0.2$$

g. Support by $(F, Location_{Midwest})$

$$Sup_{(F, Location_{Midwest})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{2,6\}|}{|\{1,2,3,4\} \cup \{2,6\}|} = \frac{1}{5} = 0.2$$

h. Support by $(F, Location_{East-coast})$

$$Sup_{(F, Location_{East-coast})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{3\}|}{|\{1,2,3,4\} \cup \{3\}|} = \frac{1}{4} = 0.25$$

i. Support by $(F, Location_{Northern})$

$$Sup_{(F, Location_{Northern})}(F, Volume_{High}) = \frac{|\{1,2,3,4\} \cap \{4\}|}{|\{1,2,3,4\} \cup \{4\}|} = \frac{1}{4} = 0.25$$

Based on the support given by each soft set, total support is computed using Definition 3. Therefore, the total support for $(F, Volume_{High})$ is given as

$$TSup(Volume_{High}) = 1 + 0 + 0.5 + 0.2 + 0.2 + 0.2 + 0.2 + 0.25 + 0.25$$

Table 4. Total support summary

Soft Set	Total Support
$(F, Volume_{High})$	2.8000
$(F, Volume_{Low})$	2.3333
$(F, Material_{Hard})$	2.1667
$(F, Material_{Medium})$	2.3667
$(F, Material_{Soft})$	2.3667
$(F, Location_{Pacific})$	2.2000
$(F, Location_{Midwest})$	2.2000
$(F, Location_{East-coast})$	1.7500
$(F, Location_{Northern})$	1.7500

Representative selection for each $(F, A_i) \in (F, A)$ is made using the formula in Definition 4. For example, representative for soft set $(F, Volume)$ is determined as follows

$$\operatorname{Re} p(F, Volume) = Max(TSup(F, Volume_{High}), TSup(F, Volume_{Low}))$$

It is clear that $(F, Volume_{High}) = 2.800$ is the representative for $(F, Volume)$ compared to $(F, Volume_{Low}) = 2.3333$. As shown in Table 4, $(F, Material)$ have two sub soft set with equal maximum value, one soft-set will be selected randomly as representative. The same step is repeated for soft-set $(F, Location)$. Thus, representatives with their value of total support for all soft set (F, A) are as follows.

$$\operatorname{Re} p(F, Volume) = (F, Volume_{High}) = 2.8000$$
$$\operatorname{Re} p(F, Material) = (F, Material_{Medium}) = 2.3667$$
$$\operatorname{Re} p(F, Location) = (F, Location_{Pacific}) = 2.200$$

Finally, super attribute representative will be determined based on Definition 5. For the above case, the decision choice will be made based on second option, since there is no repeat occurrence in representative values. Therefore, in this example, attribute Volume is selected as a decision attribute.

4 Experimental Resuts

Two simple small datasets from previous research [11] have been selected. Those dataset are Credit Card Promotion **(CCP)** and Modified Animal **(MA)**. The experiment is been done using Matlab R2008a on i5 Intel CPU with 3GB Memory. Preprocessing has been done earlier using Java.

4.1 Experiment 1 - Credit Card Promotion Dataset

The dataset as shown in Table 5, contains 10 objects with 5 categorical-value attributes, i.e. Magazine Promotion (MP), Watch Promotion (WP), Life Insurance Promotion (LIP), Credit Card Insurance (CCI) and Sex. Each attributes in the dataset contains two categories.

Table 5. Credit Card Promotion Dataset

Object	MP	WP	LIP	CCI	SEX
1	Yes	No	No	No	Male
2	Yes	Yes	Yes	No	Female
3	No	No	No	No	Male
4	Yes	Yes	Yes	Yes	Male
5	Yes	No	Yes	No	Female
6	No	No	No	No	Female
7	Yes	No	Yes	Yes	Male
8	No	Yes	No	No	Male
9	Yes	No	No	No	Male
10	Yes	Yes	Yes	No	Female

Table 6. Value of Attribute Representative of CCP Dataset

Soft Set	Value
(F,MP)	4.3389
(F,WP)	3.9722
(F,LIP)	3.8587
(F,CCI)	4.5889
(F,SEX)	4.0071

In SAR, as in Table 6, it is shown that CCI has the highest and directly selected as the decision attribute.

4.2 Experiment 2 - Modified Animal Dataset

A modified animal dataset is shown in Table 7. There are nine animals with nine attributes of categorical value namely Hair, Teeth, Eye, Feather, Feet, Eat, Milk, Fly and Swim. Six attributes namely Hair, Eye, Feather, Milk, Fly and swim have two values. One attribute namely Teeth has three values and two attributes namely Feet and Eat have four categorical values.

Table 7. Modified Animal Dataset

Animal	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Cheetah	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Giraffe	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Zebra	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Ostrich	N	N	Side	Y	Claw	Grain	N	N	N
Penguin	N	N	Side	Y	Web	Fish	N	N	Y
Albatros	N	N	Side	Y	Claw	Grain	N	Y	Y
Eagle	N	N	Forward	Y	Claw	Meat	N	Y	N
Viper	N	Pointed	Forward	N	N	Meat	N	N	N

Table 8. Value of Attribute Representative of MA Dataset

Soft Set	Total Support Value
(F,Hair)	7.8190
(F,Teeth)	7.0270
(F,Eye)	7.2738
(F,Feather)	7.3286
(F,Feet)	2.7762
(F,Eat)	3.4429
(F,Milk)	7.8190
(F,Fly)	8.2341
(F,Swim)	7.0548

As in Table 8, since there exist multi occurrence of attribute representative values, thus the first option in Definition 5 is used. Using this technique, decision can be made faster.

5 Conclusion

In this paper, Support Attribute Representative (SAR), a soft set based technique for decision making in categorical-valued information system has been proposed. In this technique, comparisons at the attribute level is firstly made before representative for each attributes and finally compared to each other. Using this technique, the execution time on selecting a decision can be reduced. The results of this research will provide useful information for decision makers to handle categorical-valued information system. Additionally, for future research, implementation and analysis of SAR using scalable dataset (high number of records, high number of attributes and high number of categories) will be carried out.

Acknowledgement

The authors would like to thanks Universiti Malaysia Terengganu for supporting this research.

References

1. Maji, P.K., Roy, A.R., Biswas, R.: An application of soft sets in a decision making problem. Computer and Mathematics with Application 44, 1077–1083 (2002)
2. Molodtsov, D.: Soft set theory-First results. Computers and Mathematics with Applications 37(4/5), 19–31 (1999)
3. Maji, P.K., Biswas, R., Roy, A.R.: Soft set theory. Computers and Mathematics with Applications 45, 555–562 (2003)
4. Ali, M.I., Feng, F., Liu, X., Min, W.K., Shabira, M.: On some new operations in soft set theory. Computers and Mathematics with Applications 57(9), 1547–1553 (2009)
5. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9(3), 589–602 (2001)
6. Kong, Z., Gao, L.Q., Wang, L.F.: Comment on "A fuzzy soft set theoretic approach to decision making problems". Journal of Computational and Applied Mathematics 223, 540–542 (2009)
7. Feng, F., Jun, Y.B., Liu, X.Y., Li, L.F.: An adjustable approach to fuzzy soft set based decision making. Journal of Computational and Applied Mathematics 234, 10–20 (2010)
8. Chen, D., Tsang, E.C.C., Yeung, D.S., Wang, X.: The parameterization reduction of soft sets and its applications. Computers and Mathematics with Applications 49(5-6), 757–763 (2005)
9. Kong, Z., Gao, L., Wang, L., Li, S.: The normal parameter reduction of soft sets and its algorithm. Computers and Mathematics with Applications 56(12), 3029–3037 (2008)
10. Herawan, T., Deris, M.M.: On Multi-soft Sets Construction in Information Systems. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS, vol. 5755, pp. 101–110. Springer, Heidelberg (2009)
11. Herawan, T., Deris, M.M.: A Framework on Rough Set-Based Partitioning Attribute Selection. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS, vol. 5755, pp. 91–100. Springer, Heidelberg (2009)

The Normal Parameter Value Reduction of Soft Sets and Its Algorithm

Xiuqin Ma and Norrozila Sulaiman

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang
26300, Kuantan, Malaysia
xueener@gmail.com,
norrozila@ump.edu.my

Abstract. Some work has been done to such issues concerning parameter reduction of soft sets. However, up to the present, few documents have focused on parameter value reduction of soft sets. In this paper, we introduce the definition of normal parameter value reduction (NPVR) of soft sets which can overcome the problem of suboptimal choice and added parameter values. More specifically, minimal normal parameter value reduction (MNPVR) is defined as a special case of NPVR and a heuristic algorithm is presented. Finally, an illustrative example is employed to show our contribution.

Keywords: Soft sets; Reduction; Parameter value reduction; Normal parameter value reduction.

1 Introduction

In recent years, there has been a rapid growth in interest in soft set theory and its applications. Soft set theory was firstly proposed by a Russian Mathematician Molodtsov [1] in 1999. It is a new mathematical tool for dealing with uncertainties, while a wide variety of theories such as probability theory, fuzzy sets [2], and rough sets [3] so on are applicable to modeling vagueness, each of which has its inherent difficulties given in [4]. In contrast to all these theories, soft set theory is free from the above limitations and has no problem of setting the membership function, which makes it very convenient and easy to apply in practice. Therefore, many applications based on soft set theory have already been demonstrated by Molodtsov [1], such as the smoothness of functions, game theory, operations research, Riemann integration, Perron integration, probability theory, and measurement theory.

Presently, great progresses of study on soft set theory have been made [5, 6, 7, 8, 9, 10, 11, 12]. And it is worthwhile to mention that some effort has been done to such issues concerning reduction of soft sets. Maji *et al.* [13] employed soft sets to solve the decision-making problem. Later, Chen *et al.* [14] pointed out that the conclusion of soft set reduction offered in [13] was incorrect, and then presented a

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 148–156, 2011.

new notion of parameterization reduction in soft sets in comparison with the definition to the related concept of attributes reduction in rough set theory. The concept of normal parameter reduction was introduced in [15], which overcome the problem of suboptimal choice and added parameter set of soft sets. An algorithm for normal parameter reduction was also presented in [15]. However, up to the present, few documents have focused on parameter value reduction of soft sets. So, in this paper, we propose a definition of normal parameter value reduction of soft sets and give a heuristic algorithm to achieve the normal parameter value reduction of soft sets.

The rest of this paper is organized as follows. Section 2 reviews the basic notions of soft set theory. Section 3 gives definitions of normal parameter value reduction and minimal normal parameter value reduction of soft sets. Furthermore we give a heuristic algorithm to achieve them, which are illustrated by an example. Finally Section 4 presents the conclusion from our study.

2 Preliminaries

In this section, we review the definition with regard to soft sets.

Let *U* be a non-empty initial universe of objects, *E* be a set of parameters in relation to objects in *U*, $P(U)$ be the power set of *U*, and $A \subset E$. The definition of soft set is given as follows.

Definition 2.1 (See [4]). *A pair* (F, A) *is called a soft set over U, where F is a mapping given by*

$$F : A \to P(U)$$

That is, a soft set over U is a parameterized family of subsets of the universe U. As an illustration, let us consider the following example, which is quoted directly from [4].

Example 2.1. Let be a soft set (F, E) describe the "attractiveness of houses" that Mr. X is going to purchase. Suppose that $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ and $E = \{e_1, e_2, e_3, e_4, e_5\}$, where there are six houses in the universe *U* and *E* is a set of parameters, e_i $(i = 1,2,3,4,5)$ standing for the parameters "expensive", "beautiful", "wooden", "cheap", and "in the green surroundings" respectively.

Suppose that we have

$F(e_1) = \{h_2, h_4\}$, $F(e_2) = \{h_1, h_3\}$, $F(e_3) = \phi$, $F(e_4) = \{h_1, h_3, h_5\}$, and $F(e_5) = \{h_1\}$, where $F(e_i)$ means a subset of U which elements match the parameter e_i. Then we can view the soft set (F,E) as consisting of the following collection of approximations:

$$(F,E)=\left\{\begin{array}{c}\text{expensive houses}=\{h_2,h_4\},\\ \text{beautiful houses}=\{h_1,h_3\}\\ \text{wooden houses}=\phi\\ \text{cheap houses}=\{h_1,h_3,h_5\}\\ \text{in the green surrounding houses}=\{h_1\}\end{array}\right\}$$

Each approximation has two parts, a predicate p and an approximate value set v. For example, for the approximation "expensive houses $=\{h_2,h_4\}$", we have the predicate name of expensive houses and the approximate value set or value set is $\{h_2,h_4\}$.Thus, a soft set (F,E) can be viewed as a collection of approximations below:

$$(F,E)=\{p_1=v_1,p_2=v_2,p_3=v_3,\cdots,p_n=v_n\}.$$

The soft set is a mapping from parameter to the crisp subset of universe. From such case, we may see the structure of a soft set can classify the objects into two classes (yes/1 or no/0). Thus we can make a one-to-one correspondence between a Boolean-valued information system and a soft set, as stated in Proposition 2.1.

Definition 2.2. *An information system is a 4-tuple (quadruple)* $S=(U,A,V,f)$, *where* $U=\{u_1,u_2,\cdots,u_{|U|}\}$ *is a non-empty finite set of objects,* $A=\{a_1,a_2,\cdots,a_{|A|}\}$ *is a non-empty finite set of attributes,* $V=\bigcup_{a\in A}V_a$, V_a *is the domain (value set) of attribute a,* $f:U\times A\to V$ *is an information function such that* $f(u,a)\in V_a$, *for every* $(u,a)\in U\times A$, *called information (knowledge) function.*

An information system is also called a knowledge representation system or an attribute-valued system and can be intuitively expressed in terms of an information table. In an information system $S=(U,A,V,f)$, if $V_a=\{0,1\}$, for every $a\in A$, then S is called a Boolean-valued information system.

Proposition 2.1. *If* (F,E) *is a soft set over the universe U, then* (F,E) *is a Boolean-valued information system* $S=(U,A,V_{\{0,1\}},f)$.

From Proposition 2.1, a soft set (F,E) as in Example 2.1, it can be represented as a Boolean table as follows:

Table 1. Tabular representation of a soft set in the above example

U	*e₁*	*e₂*	*e₃*	*e₄*	*e₅*
h_1	0	1	0	1	1
h_2	1	0	0	0	0
h_3	0	1	1	1	0
h_4	1	0	1	0	0
h_5	0	0	1	1	0
h_6	0	0	0	0	0

3 The Normal Parameter Value Reduction of Soft Sets

In this section, we depict a definition of normal parameter value reduction of soft sets (NPVR). Furthermore, we introduce the minimal normal parameter value reduction (MNPVR) of soft sets as a special case of NPVR and then present a heuristic algorithm to achieve it. Finally, an illustrative example is given.

3.1 Definition of Normal Parameter Value Reduction of Soft Sets

Suppose $U = \{h_1, h_2, \cdots, h_n\}$, $E = \{e_1, e_2, \cdots, e_m\}$, (F, E) is a soft set with tabular representation. Define $f_E(h_i) = \sum_j h_{ij}$, where h_{ij} are the entries in the table of (F, E).

Definition 3.1 (See [15]). *With every subset of parameters* $B \subseteq A$ *, an indiscernibility relation* $IND(B)$ *is defined by*

$$IND(B) = \{(h_i, h_j) \in U \times U : f_B(h_i) = f_B(h_j)\}.$$

For soft sets (F, E), $U = \{h_1, h_2, \cdots, h_n\}$, the decision partition is referred to as

$$C_E = \{\{h_1, h_2, ..., h_i\}_{f_1}, \{h_{i+1}, ..., h_j\}_{f_2}, ..., \{h_k, ..., h_n\}_{f_s}\},$$

where for subclass $\{h_v, h_{v+1}, ..., h_{v+w}\}_{f_i}, f_E(h_v) = f_E(h_{v+1}) = ... = f_E(h_{v+w}) = f_i$, and $f_1 \geq f_2 \geq ... \geq f_s$, s is the number of subclasses. In other words, objects in U are classified and ranked according to value of $f_E(.)$ based on the indiscernibility relation.

Definition 3.2. *Given a Boolean-valued information system* $S = \left(U, A, V_{\{0,1\}}, f\right)$ *corresponding to a soft set* (F, E) *, we define:*

$$(1)(e_j, h_{ij}) \overset{def}{=} \{h_i \mid (h_i \in U) \wedge (f_{e_j}(h_i) = h_{ij}), i = 1,2,...,n; j = 1,2,...,m\};$$

$$(2)((e_1, h_{i1}), (e_2, h_{i2}),..., (e_m, h_{im})) \overset{def}{=} \{h_i \mid (h_i \in U) \wedge (f_{e_1}(h_i) = h_{i1,} f_{e_2}(h_i) = h_{i2,} ...,$$
$$f_{e_m}(h_i) = h_{im,} i = 1,2,...,n\};$$

$$(3)H_E(h_i) = \underset{j \in E}{\cup}(e_j, h_{ij}) = \{(e_1, h_{i1}), (e_2, h_{i2}),..., (e_m, h_{im})\}, i = 1,2,...,n.$$

Definition 3.3. *Denote* $A_1, A_2,..., A_n \subset E$ *as subsets, if there exist subsets* $A_1, A_2,..., A_n$ *satisfying* $f_{A_1}(h_1) = f_{A_2}(h_2) = \cdots = f_{A_n}(h_n) = t (t \le f_s)$ *, then the parameter values* $H_{A_i}(h_i)(i = 1,...,n)$ *are dispensable, otherwise, they are indispensable.* $H_{B_i}(h_i)(i = 1,...,n)$ *is defined as a normal parameter value reduction, if the two conditions as follows are satisfied*

(1) $H_{B_i}(h_i)(i = 1,...,n)$ *is indispensable*

(2) $f_{E-B_1}(h_1) = f_{E-B_2}(h_2) = ... = f_{E-B_n}(h_n) = t$

From the definition of normal parameter value reduction, we know normal parameter value reduction of soft sets keeps the classification ability and rank invariant for decision making. That is, after reducing dispensable parameter value, the decision partition is $C_E' = \left\{\{h_1, h_2,..., h_i\}_{f_1 - t}, \{h_{i+1},..., h_j\}_{f_2 - t},..., \{h_k,..., h_n\}_{f_s - t}\right\}$

Definition 3.4. *Denote* $A_1, A_2,..., A_n \subset E$ *as subsets, if there exist subsets* $A_1, A_2,..., A_n$ *satisfying* $f_{A_1}(h_1) = f_{A_2}(h_2) = \cdots = f_{A_n}(h_n) = t$ *, then the parameter values* $H_{A_i}(h_i)(i = 1,...,n)$ *are dispensable, otherwise, they are indispensable.* $H_{B_i}(h_i)(i = 1,...,n)$ *are defined as a minimal normal parameter value reduction, if the three conditions as follows are satisfied*

(1) $H_{B_i}(h_i)(i = 1,...,n)$ *is indispensable*

(2) $f_{E-B_1}(h_1) = f_{E-B_2}(h_2) = ... = f_{E-B_n}(h_n) = t$

(3) $t = f_s$ *(* f_s *is the minimum decision choice value)*

From the above definition it follows that MNPVR does not differ essentially from NPVR. Obviously, MNPVR is a special case of NPVR. Intuitively speaking, MNPVR leads to the final decision partition $C_E' = \{\{h_1, h_2, ..., h_i\}_{f_1 - t}, \{h_{i+1}, ..., h_j\}_{f_2 - t}, ..., \{h_k, ..., h_n\}_{f_s - t = 0}\}$.

3.2 Algorithm of Minimal Normal Parameter Value Reduction of Soft Sets

Here below, we provide an algorithm to illustrate how to achieve the minimal normal parameter value reduction of soft sets.

(1) Input the soft set (F, E) and the parameter set E;

(2) Delete the parameter values denoted by 0.

(3) If $f_s = t \neq 0$, reduce the t parameter values denoted by 1 for every $h_i (h_i \in U, 0 \leq i \leq n)$ until $f_s = 0$.

(4) Put the remainder values as the minimal normal parameter value reduction which satisfies $C_E' = \{\{h_1, h_2, ..., h_i\}_{f_1 - t}, \{h_{i+1}, ..., h_j\}_{f_2 - t}, ..., \{h_k, ..., h_n\}_{f_s - t = 0}\}$.

3.3 Example

Example 3.1. Let (F, E) be a soft set with the tabular representation displayed in Table 2. Suppose that $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$, and $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$

Table 2. A soft set (F, E)

h	*e₁*	*e₂*	*e₃*	*e₄*	*e₅*	*e₆*	*e₇*	*e₈*	*f(.)*
h_1	1	0	0	0	0	1	0	1	3
h_2	0	1	0	1	1	1	0	1	5
h_3	1	0	1	1	1	1	0	1	6
h_4	1	0	0	1	0	0	0	1	3
h_5	0	1	0	0	1	0	0	1	3
h_6	1	0	0	1	1	0	0	1	4

From table 2, the decision partition is $C_E = \{\{h_3\}_6, \{h_2\}_5, \{h_6\}_4, \{h_1, h_4, h_5\}_3\}$. Clearly $f_E(h_3) = 6$ is the maximum choice value, thus h_3 is the optimal choice object. h_2 is the suboptimal choice object.

Table 3. A normal parameter value reduction table of original table (Table 2)

h	*e₁*	*e₂*	*e₃*	*e₄*	*e₅*	*e₆*	*e₇*	*e₈*	*f(.)*
h_1	-	-	-	-	-	1	-	1	2
h_2	-	-	-	1	1	1	-	1	4
h_3	-	-	1	1	1	1	-	1	5
h_4	-	-	-	1	-	-	-	1	2
h_5	-	-	-	-	1	-	-	1	2
h_6	-	-	-	1	1	-	-	1	3

Given t=1 (namely, $f_{A_1}(h_1)=f_{A_2}(h_2)=\cdots=f_{A_n}(h_n)=t=1$), it is evident that, $A_1=\{e_1,e_2,e_3,e_4,e_5,e_7\}, A_2=\{e_1,e_2,e_3,e_7\}, A_3=\{e_1,e_2,e_7\}, A_4=\{e_1,e_2,e_3,e_5,e_6,e_7\}, A_5=$ $\{e_1,e_2,e_3,e_4,e_6,e_7\}, A_6=\{e_1,e_2,e_3,e_6,e_7\}$in this example. The normal parameter value reduction of $(F,E)$is clearly shown in Table 3. And then we get that the decision partition is $C_E=\{\{h_3\}_5,\{h_2\}_4,\{h_6\}_3,\{h_1,h_4,h_5\}_2\}$. h_3 is still the optimal choice object. The normal parameter value reduction keeps the classification ability and rank invariant for decision making.

Table 4. A minimal normal parameter value reduction table of original table (Table 2)

h	*e₁*	*e₂*	*e₃*	*e₄*	*e₅*	*e₆*	*e₇*	*e₈*	*f(.)*
h_1	-	-	-	-	-	-	-	-	0
h_2	-	-	-	-	-	1	-	1	2
h_3	-	-	-	-	1	1	0	1	3
h_4	-	-	-	-	-	-	-	-	0
h_5	-	-	-	-	-	-	-	-	0
h_6	-	-	-	-	-	-	-	1	1

Clearly $f_s = 3$ is the minimum decision choice value in original Table 2. Thus, let $f_{A_1}(h_1) = f_{A_2}(h_2) = \cdots = f_{A_n}(h_n) = f_s = t = 3$. And then we can obtain

$$A_1 = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}, A_2 = \{e_1, e_2, e_3, e_4, e_5, e_7\}, A_3 = \{e_1, e_2, e_3, e_4, e_7\},$$
$$A_4 = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}, A_5 = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}, A_6 = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}.$$

The final decision partition is $C_E = \{\{h_3\}_3, \{h_2\}_2, \{h_6\}_1, \{h_1, h_4, h_5\}_0\}$. The results from Table 3 and Table 4 indicate that MNPVR can delete more parameter values in comparison with NPVR, in the case of keeping the classification ability and rank invariant for decision making. Thus MNPVR can be generally interpreted as the minimal degree of NPVR.

4 Conclusion

Some work on parameter reduction of soft sets has been done. They presented a new notion of parameterization reduction in soft sets in comparison with the definition to the related concept of attributes reduction in rough set theory and the concept of normal parameter reduction which overcome the problem of suboptimal choice and added parameter set of soft sets. Unfortunately, the two algorithms are complicated and time-consuming. And up to the present, few documents have focused on parameter value reduction of soft sets. So, in this paper, we propose a definition of normal parameter value reduction (NPVR) of soft sets. More specifically, minimal normal parameter value reduction (MNPVR) is defined as a special case of NPVR and a heuristic algorithm is presented to achieve it, which is very easy to understand and carried out. Finally, an illustrative example is employed to show our contribution.

References

1. Molodtsov, D.: Soft set theory_First results. Computers and Mathematics with Applications 37(4/5), 19–31 (1999)
2. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
3. Pawlak, Z.: Rough sets. International Journal Information Computer Science 11, 341–356 (1982)
4. Molodtsov, D.: The Theory of Soft Sets. URSS Publishers, Moscow (2004) (in Russian)
5. Pei, D., Miao, D.: From soft sets to information systems. In: The proceeding of 2005 IEEE International Conference on Granular Computing, IEEE GrC 2005, pp. 617–621. IEEE Press, Los Alamitos (2005)
6. Herawan, T., Mat Deris, M.: A direct proof of every rough set is a soft set. In: Proceeding of the Third Asia International Conference on Modeling and Simulation, pp. 119–124. AMS, Bali (2009)
7. Feng, F., Li, C., Davvaz, B., Ali, M.I.: Soft sets combined with fuzzy sets and rough sets: A tentative approach. In: Soft Computing - A Fusion of Foundations, Methodologies and Applications, pp. 899–911. Springer, Heidelberg (2009)
8. Feng, F.: Generalized rough fuzzy sets based on soft sets. In: The Proceeding of 2009 International Workshop on Intelligent Systems and Applications, ISA 2009, Wuhan, China, pp. 1–4 (2009)

9. Ali, M.I., Feng, F., Liu, X., Min, W.K., Shabira, M.: On some new operations in soft set theory. Comput. Math. Appl. 57(9), 1547–1553 (2009)
10. Zou, Y., Xiao, Z.: Data analysis approaches of soft sets under incomplete information. Knowl.-Based Syst. 21(8), 941–945 (2008)
11. Xiao, Z., Gong, K., Xia, S., Zou, Y.: Exclusive disjunctive soft sets. Computers and Mathematics with Applications 59(6), 2128–2137 (2010)
12. Herawan, T., Mat Deris, M.: A Soft Set Approach for Association Rules Mining. Knowledge Based Systems (2010), doi:10.1016/j.knosys.2010.08.005
13. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
14. Chen, D., Tsang, E.C.C., Yeung, D.S., Wang, X.: The parameterization reduction of soft sets and its applications. Computers and Mathematics with Applications 49(5–6), 757–763 (2005)
15. Kong, Z., Gao, L., Wang, L., Li, S.: The normal parameter reduction of soft sets and its algorithm. Computers and Mathematics with Applications 56(12), 3029–3037 (2008)

Improving Language Identification of Web Page Using Optimum Profile

Choon-Ching Ng[1,⋆] and Ali Selamat[2]

[1] Faculty of Computer Systems & Software Engineering,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Gambang, Kuantan Pahang, Malaysia
chooncching@ump.edu.my

[2] Faculty of Computer Science & Information Systems,
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
aselamat@utm.my

Abstract. Language is an indispensable tool for human communication, and presently, the language that dominates the Internet is English. Language identification is the process of determining a predetermined language automatically from a given content (e.g., English, Malay, Danish, Estonian, Czech, Slovak, etc.). The ability to identify other languages in relation to English is highly desirable. It is the goal of this research to improve the method used to achieve this end. Three methods have been studied in this research are distance measurement, Boolean method, and the proposed method, namely, optimum profile. From the initial experiments, we have found that, distance measurement and Boolean method is not reliable in the European web page identification. Therefore, we propose optimum profile which is using N-grams frequency and N-grams position to do web page language identification. The result show that the proposed method gives the highest performance with accuracy 91.52%.

Keywords: N-grams profile, rank-order statistics, distance measurement, Boolean method, optimum profile.

1 Introduction

Language identification is used frequently in a number of applications, such as machine translation, information retrieval, speech recognition, and text categorization. Among researches of text-based language identification, N-grams is perhaps the most widely used and studied [1]. The N-grams method, which is a sub-sequence of N objects from a longer sequence, when rank-order statistics on N-grams profile are adopted and the distance measurement is used to identify the predefined language of a particular content.

It is being argued that text-based language identification is a completely solved problem. However, we have found that improvements are still needed

⋆ Corresponding author.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 157–166, 2011.

because of several problems arise when dealing with web page language identification. Firstly, the web pages contain multiple languages which may produce faulty output in related language systems. Secondly, web page language identification is difficult due to plethora of internation terms and proper names occuring in the internet. Other issues are web page format, encoding, spelling and grammar errors [2,3].

This paper is organized as follows: Related works on language identification is described in Section 2. Next, data preparation and language identification using the distance measurement, Boolean method, and optimum profile are explained in Section 3. The experimental results based on confusion matrix and accuracy are detailed out in Section 4. The conclusion of the research is given in Section 5.

2 Related Works

Human usually don't have any need for language identifiers, however the field of human language technology covers a number of research activities, such as the coding, identification, interpretation, translation and generation of language. The aim of such research is to enable humans to communicate with machines using natural language skills. Language technology research involves many disciplines, such as linguistics, psychology, electrical engineering and computer science. Cooperation among these disciplines is needed to create multimodal and multimedia systems that use the combination of text, speech, facial cues and gestures, both to improve language understanding and to produce more natural language processing by animated characters [4,5].

Language technologies play a key role in the age of information [5]. Today, almost all device systems combine language understanding and generation that allow people to interact with computers using text or speech to obtain information, to study, to do business, and to communicate with each other effectively [6]. The technology convergence in the processing of text, speech, and images has lead to the particular ability to make sense of the massive amounts of information now available via computer networks. For example, if a student wants to gather information about the art of getting things done, he or she can set in motion a set of procedures that locate, organize, and summarize all available information related to the topic from books, periodicals, newspapers, and so on. Translation of texts or speech from one language to another is needed to access and interpret all available material and present it to the student in his or her native language. As a result, it increases academic interests of the student [6,7].

Some works have been reported to detect the language of a particular web page. They are decision tree neural networks [3], discrete HMMs [2], short letter sequences (N-grams) [8], and metadata description [9]. A variety of features have been used for language identification. These includes: the presence of particular characters [10,11], written words [12,13], and N-grams [1,14].

3 Method

In this section, we describe the distance measurement, Boolean method, and optimum profile as shown in Figure 1. First of all, data sets of languages have been collected from news website. Then, these data sets were saved in unicode form by setting the file name corresponding to the target language. Many types of encoding have been used on the web document to ensure that character processing is not miscalculated. We have converted the identified encoding into the unicode encoding as the latter is able to accommodate all encoding types by the use of specific numeric number. In this work, we have collected European web pages as experimental data sets such as Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish, and the corresponding annotation are bul, cze, dan, dut, eng, est, fin, fre, ger, gre, hun, iri, ita, lat, lit, mat, pol, por, rom, slo, sle, spa, and swe, respectively. Each language consists of 1000 web pages, 100 units were used for training, 500 units were randomly selected from the remaining data sets as testing data. Threshold has been set to top 100 units of language features. N-grams were mixed by unigram, bigrams, and trigrams, but statistical anaysis was done independently.

Feature selection determines the appropriate features or attributes to be used in language identification. It is based on N-grams frequency (NF) and rank-order statistics [15]. For NF, it is based on the occurences of the particular N-grams (ngm) in a document, not the whole data set. The number of a particular ngm contributed to the document is an important factor in a language identification. For example, the ngm 'ber' appears in Malay more frequently than in English, so a Malay document has higher occurences of that ngm. The formula of NF_L is given by equation 1, where T_d is the total N-grams in that

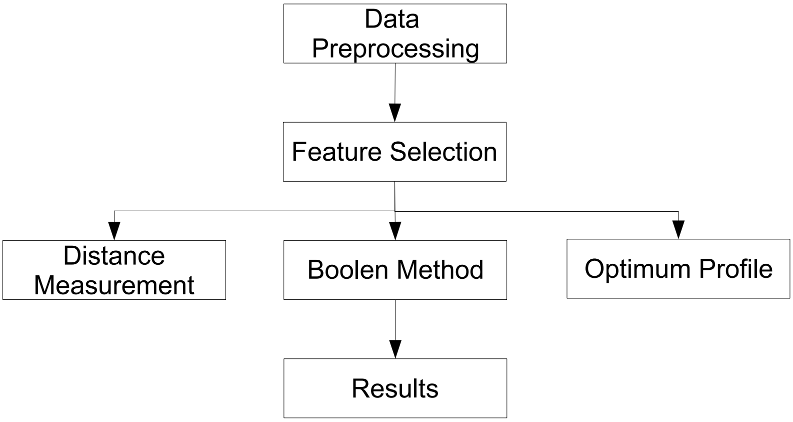


Fig. 1. Research framework of proposed method

document d, L is the target language, and D is the data set. Each unigram, bigrams, and trigrams calculation is done separately and also the highest NF is selected as features. Finally, ngm is sorted based on rank-order statistics. Details of experimental setup and measurement have been described in the paper of Selamat and Ng [3,16,17].

$$NF_L\left(ngm\right) = \sum_{d=1}^{D}\left(\frac{\sum ngm_d}{T_d}\right) \tag{1}$$

3.1 Distance Measurement

Distance measurement has been proposed by Cavnar and Trenkle [1], they have used rank-order statistics on N-grams profiles in order to find out closest profile as winner of language identification and text categorization. Figure 2 illustrates the distance measurement of N-grams profile. First, training profile of each language is generated from the desired data set by using N-grams frequency. After

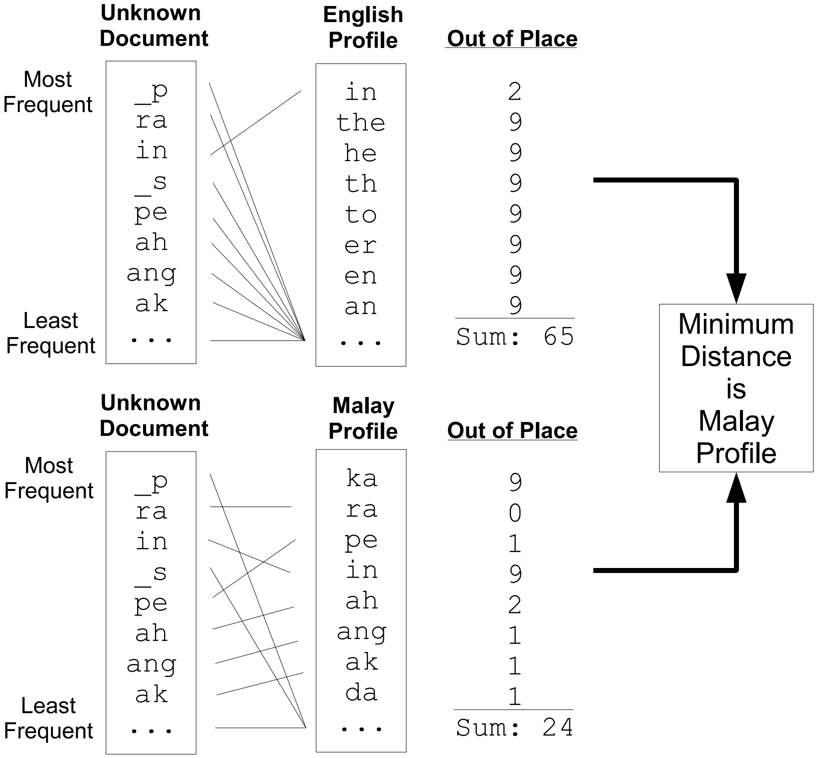


Fig. 2. Distance measurement of N-grams profile (note: it is assumed that for those not found N-grams in this figure is assigned with a maximum value nine) [1]

that, rank-order statistics are applied on the language profile to sort the N-grams from most frequent to least frequent. Same process goes on unknown document or target document. Then, unknown document profile is compared with all profiles of desired languages. Out of place is the distance between desired N-grams and target N-grams. Finally, minimum distance of one particular profile is selected as winner based on the sum out of place.

3.2 Boolean Method

Boolean method has been used to measure matching rates between target profile and training profile. It is different with distance measurement which is depends on N-grams frequency. Instead, this method returns value of one if one particular N-gram from the target profile is found on the desired profile. Otherwise, it returns value of zero if there is no match. After that, matching rate is derived by dividing the total Boolean value to total number of distinct N-grams in the target profile. Finally, the maximum matching rate is selected as winner among the training profiles.

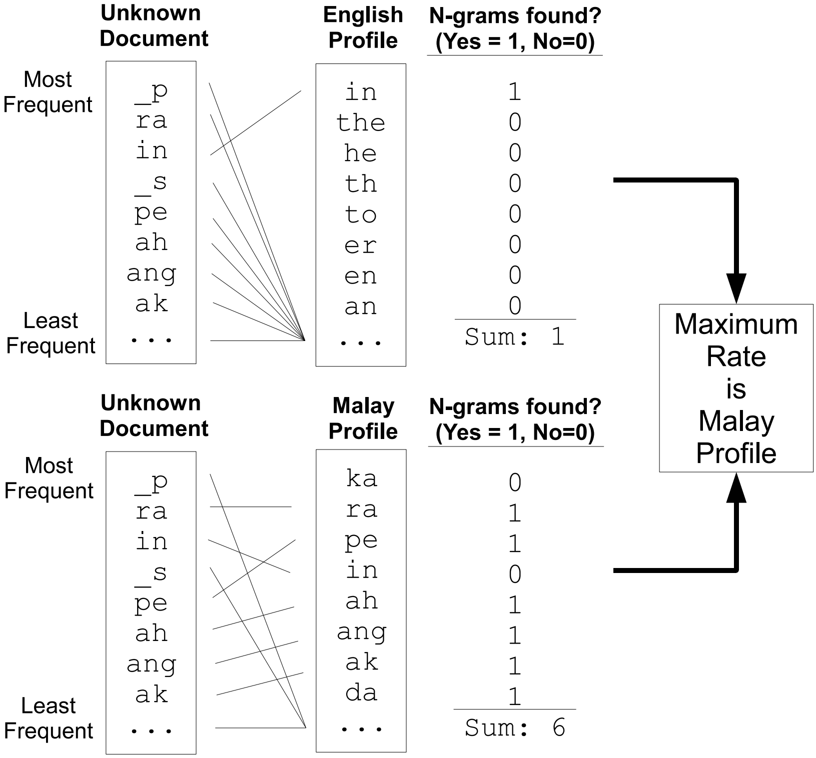
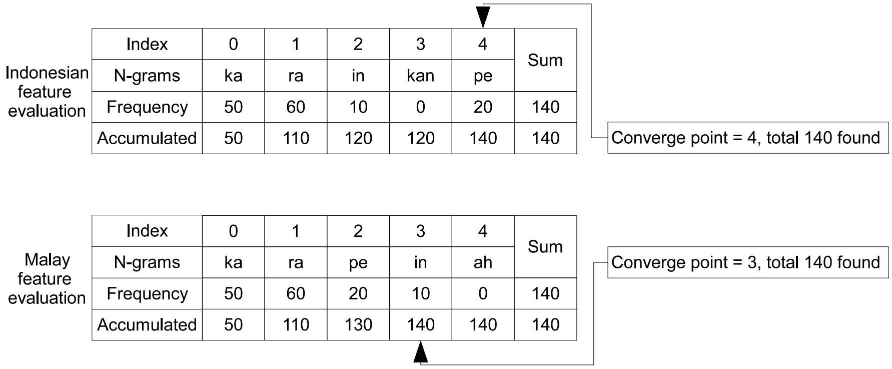


Fig. 3. Boolean method of N-grams profile (note: it is assumed that for those not found N-grams in this figure is assigned with a zero value) [18]

3.3 Optimum Profile

Figure 4 shows the example of proposed method N-grams optimum profile. This method makes use of N-grams frequency and N-grams position. Accumulated N-grams frequencies is the first identifier of language identification and it is followed by converge point which is to determine the fastest convergence of N-grams position. A random double is added to converge point to increase the level of discriminant. Indonesian N-grams profile consists of 'ka', 'ra', 'in', 'kan', and 'pe'; however Malay N-grams profile is comprised of 'ka', 'ra', 'pe', 'in', and 'ah'. Each N-grams frequency of Indonesian is 50, 60, 10, 0, 20, and Malay is 50, 60, 20, 10, and 0; while the accumulated frequencies are 50, 110, 120, 120, 140, 50, 110, 130, 140, and 140, respectively. Converge point of Indonesian and Malay are 4 and 3, respectively. Winner of this example is Malay due to the converge point is smaller than Indonesian.



Indonesian feature evaluation

Index	0	1	2	3	4	Sum
N-grams	ka	ra	in	kan	pe	
Frequency	50	60	10	0	20	140
Accumulated	50	110	120	120	140	140

Converge point = 4, total 140 found

Malay feature evaluation

Index	0	1	2	3	4	Sum
N-grams	ka	ra	pe	in	ah	
Frequency	50	60	20	10	0	140
Accumulated	50	110	130	140	140	140

Converge point = 3, total 140 found

Fig. 4. Proposed N-grams Optimum Profile

4 Experimental Results

In this following subsections, we discuss the confusion matrix and accuracy of European web page language identification. Three methods have been evaluated which are distance measurement, Boolean method, and optimum profile. European data sets have been used in the experiment with total 23 languages. Threshold was set to top 100 features of each language.

4.1 Confusion Matrix of Web Page Language Identification

Table 1 shows the confusion matrix of European web page language identification using distance measurement. It is observed that the Bulgarian, Czech, Estonian, and Greek give worst results with correctly predicted samples are 0, 66, 1, and 4, respectively. Other languages more than 321 samples were correctly predicted. Finnish and Hungarian have achieved the best identification results which are 100% correctness.

Table 1. Confusion matrix of distance measurement on European web page language identification

		Predicted Language																						
		bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
Desired Language	bul	0	0	0	0	0	8	0	0	0	377	0	115	0	0	0	0	0	0	0	0	0	0	0
	cze	0	66	0	0	0	7	0	0	0	3	0	0	0	0	0	0	16	0	0	163	245	0	0
	dan	1	0	471	1	0	0	0	0	2	11	4	0	0	0	0	0	0	0	0	0	0	0	10
	dut	0	0	0	396	0	46	0	0	0	53	2	0	0	0	0	0	2	0	0	0	0	0	1
	eng	7	0	0	0	473	0	0	0	0	18	2	0	0	0	0	0	0	0	0	0	0	0	0
	est	8	3	23	1	4	1	21	2	0	233	117	5	0	6	1	29	36	6	2	0	1	0	1
	fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fre	0	0	0	0	0	0	0	495	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	ger	0	0	0	0	0	0	0	0	499	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	gre	55	12	3	1	0	59	219	1	6	4	3	31	2	22	0	38	9	0	2	14	18	0	1
	hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
	iri	1	4	0	0	0	95	0	0	0	2	0	391	0	1	0	1	0	0	0	1	2	0	2
	ita	17	1	0	0	0	26	1	0	0	0	21	0	422	2	0	0	3	0	0	3	3	0	1
	lat	20	13	0	1	1	36	0	3	0	42	4	13	0	347	2	3	4	0	0	9	0	0	2
	lit	0	0	0	0	0	0	0	0	0	1	0	0	0	3	494	0	1	0	0	1	0	0	0
	mat	3	16	0	0	0	14	1	1	0	27	5	1	0	0	0	432	0	0	0	0	0	0	0
	pol	2	3	0	0	0	34	0	0	0	34	1	0	0	0	0	0	417	0	0	9	0	0	0
	por	8	0	0	0	0	6	0	0	0	15	1	1	1	0	0	0	4	463	0	0	0	1	0
	rom	2	1	0	1	1	15	2	1	0	9	7	0	12	0	0	2	4	0	442	0	0	0	1
	slo	6	17	0	2	0	59	0	0	1	77	0	0	0	0	1	0	13	0	0	321	3	0	0
	sle	0	0	0	0	0	1	0	0	0	2	0	1	0	0	0	0	25	0	0	49	422	0	0
	spa	1	0	0	0	14	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	482	0
	swe	9	4	1	1	0	12	0	1	2	82	21	0	0	0	0	5	6	0	0	1	0	0	355

Table 2. Confusion matrix of Boolean method on European web page language identification

		Predicted Language																						
		bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
Desired Language	bul	180	0	0	0	320	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	cze	0	45	0	0	1	0	0	0	0	0	0	0	16	0	0	0	0	0	5	0	432	1	0
	dan	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dut	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	eng	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	est	0	0	21	2	17	11	247	0	0	0	0	0	119	0	62	0	0	0	8	0	0	0	13
	fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fre	0	0	0	0	1	0	0	499	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ger	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	gre	2	0	0	0	4	0	0	0	0	494	0	0	0	0	0	0	0	0	0	0	0	0	0
	hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
	iri	0	0	0	0	125	0	0	0	0	0	0	375	0	0	0	0	0	0	0	0	0	0	0
	ita	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0
	lat	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0
	lit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0
	mat	0	0	1	0	4	0	0	0	0	0	0	0	10	0	0	485	0	0	0	0	0	0	0
	pol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0
	por	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	498	0	0	0	0	0
	rom	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	498	0	0	0	0
	slo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	499	0	0	0
	sle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0
	spa	0	0	0	0	302	0	0	2	0	0	1	0	1	0	0	0	0	0	1	0	0	193	0
	swe	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	499

Table 2 illustrates the confusion matrix of European web page language identification using Boolean method. Bulgarian, Czech, Estonian, and Spanish have achieved accuracy of identification below 50%. Total correctly predicted samples are 180, 45, 11, and 193, respectively. Irish has been predicted as English with 125 samples and the remaining are correct samples. Other languages give good results with more than or equal 485 correct samples. It is slightly better than distance measurement.

Table 3 depicts the results of European web page language identification by using optimum profile. It is noticed that the performance of identification has been increased with only two languages give worst results. They are language of Czech and Estonian. Only 45 samples of Czech and 9 samples of Estonian have been correctly predicted. The remaining 21 languages have correctly predicted more than 490 samples.

Table 3. Confusion matrix of optimum profile on European web page language identification

		Predicted Language																						
		bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
Desired Language	bul	496	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	cze	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	427	28	0	0
	dan	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dut	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	eng	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	est	0	0	111	1	1	9	324	0	0	0	0	0	0	2	23	0	0	0	0	0	1	0	28
	fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fre	0	0	0	0	0	0	0	498	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
	ger	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	gre	1	0	0	0	2	0	0	0	2	491	0	0	0	1	1	0	0	1	0	0	0	1	0
	hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
	iri	0	0	0	0	0	0	0	0	0	0	0	496	4	0	0	0	0	0	0	0	0	0	0
	ita	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0
	lat	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0
	lit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0
	mat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0
	pol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0
	por	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	499	0	0	0	1	0
	rom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0
	slo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0
	sle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0
	spa	0	1	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	492	0
	swe	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	499

4.2 Accuracy of Web Page Language Identification

Figure 5 shows the overall accuracy of European web page language identification. It is found that the accuracy of distance measurement, Boolean method, and optimum profile is 72.98%, 85.01%, and 91.52%. Distance measurement is having dimension problem in which similar distance might be found in more than one language and usually the smallest one is frequently encountered. Boolean

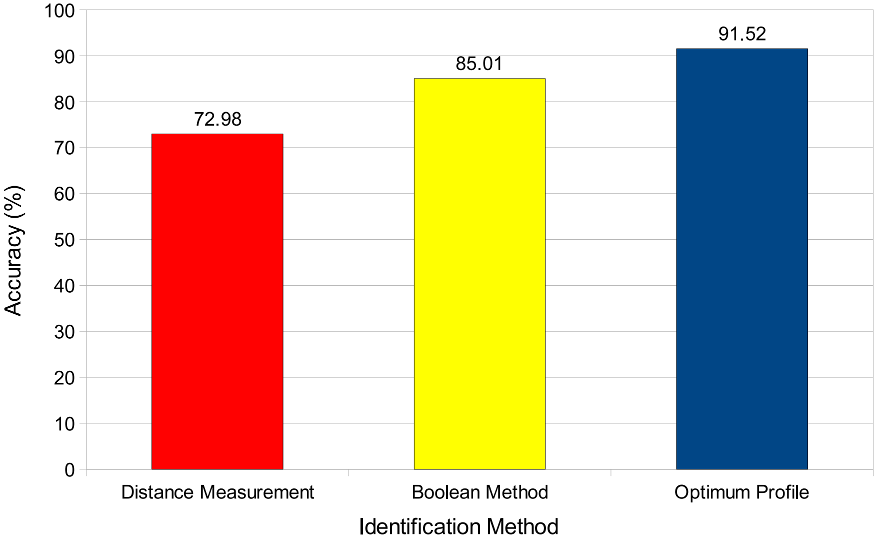


Fig. 5. Overall accuracy of European web page language identification

method is not reliable if two or more languages appear same N-grams frequency. Therefore, it has been proved that optimum profile gives the best performance in European web page language identification.

5 Conclusion

Language identification is important in a vast variety of natural language processing systems. If we are to trust an information retrieval system to classify documents with little or no human oversight, we require a system that is capable of operating at a high level of high accuracy. Therefore, we have proposed optimum profile to cope with the limitations found in both distance measurement and Boolean method. From the experiments, it is concluded that optimum profile performs better than others. In the future works, the issues of multilingual web page, noise tolerent of language identifier, minority languages, and data dimensionality will be investigated.

Acknowledgments. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Center, Universiti Malaysia Pahang (UMP) and Universiti Teknologi Malaysia (UTM). The authors are also grateful to the anonymous reviewers for their valuable and insightful comments.

References

1. Cavnar, W., Trenkle, J.: N-gram-based text categorization. In: Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, USA, pp. 161–175 (1994)
2. Xafopoulos, A., Kotropoulos, C., Almpanidis, G., Pitas, I.: Language identification in web documents using discrete hmms. Pattern Recognition 37(3), 583–594 (2004)
3. Selamat, A., Ng, C.: Arabic script web page language identifications using decision tree neural networks. Pattern Recognition 44(1), 133–144 (2011)
4. Muthusamy, Y., Spitz, A.: Automatic language identification. In: Cole, R., Mariani, J., Uszkoreit, H., Varile, G., Zaenen, A., Zampolli, A. (eds.) Survey of the State of the Art in Human Language Technology, pp. 255–258. Cambridge University Press, Cambridge (1997)
5. Constable, P., Simons, G.: Language identification and it: Addressing problems of linguistic diversity on a global scale. In: Proceedings of the 17th International Unicode Conference, SIL Electronic Working Papers, San José, California, pp. 1–22 (2000)
6. Abd Rozan, M.Z., Mikami, Y., Abu Bakar, A.Z., Vikas, O.: Multilingual ict education: Language observatory as a monitoring instrument. In: Proceedings of the South East Asia Regional Computer Confederation 2005: ICT Building Bridges Conference, Sydney, Australia, vol. 46 (2005)
7. McNamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Information Retrieval 7(1), 73–97 (2004)
8. Martins, B., Silva, M.J.: Language identification in web pages. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 764–768 (2005)

9. Simons, G.F.: Language identification in metadata descriptions of language archive holdings. In: Workshop on Web-Based Language Documentation and Description, Philadelphia, USA (2000)
10. Hakkinen, J., Tian, J.: N-gram and decision tree based language identification for written words. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 335–338 (2001)
11. Takcı, H., Soğukpınar, İ.: Letter Based Text Scoring Method for Language Identification. In: Yakhno, T. (ed.) ADVIS 2004. LNCS, vol. 3261, pp. 283–290. Springer, Heidelberg (2004)
12. Biemann, C., Teresniak, S.: Disentangling from babylonian confusion – unsupervised language identification. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 773–784. Springer, Heidelberg (2005)
13. Hammarstrom, H.: A fine-grained model for language identification. In: Workshop of Improving Non English Web Searching, Amsterdam, The Netherlands, pp. 14–20 (2007)
14. da Silva, J.F., Lopes, G.P.: Identification of document language is not yet a completely solved problem. In: Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce, pp. 212–219. IEEE Computer Society, Washington, DC, USA (2006)
15. Ng, C., Selamat, A.: Improve feature selection method of web page language identification using fuzzy artmap. International Journal of Intelligent Information and Database Systems 4(6), 629–642 (2010)
16. Selamat, A., Subroto, I., Ng, C.: Arabic script web page language identification using hybrid-knn method. International Journal of Computational Intelligence and Applications 8(3), 315–343 (2009)
17. Selamat, A., Ng, C.: Arabic script language identification using letter frequency neural networks. International Journal of Web Information Systems 4(4), 484–500 (2008)
18. Choong, C.Y., Mikami, Y., Marasinghe, C.A., Nandasara, S.T.: Optimizing n-gram order of an n-gram based language identification algorithm for 68 written languages. International Journal on Advances in ICT for Emerging Regions 2(2), 21–28 (2009)

The Grounded Process of KMS Adoption: The Case Study of Knowledge Sharing and CoPs in Oil and Gas Industry in Malaysia

Sureena Matayong and Ahmad Kamil Bin Mahmood

Department of Computer and Information Sciences
Universiti Teknologi Petronas, Malaysia
smatayong@gmail.com, kamilmh@petronas.com.my

Abstract. This paper presents the adoption of a knowledge management system (KMS). It is an innovation in the field of IT and its adoption rests within the literature of IT adoption theories. This research documents via Grounded Theory (GT), the adoption of KMS at an Oil and Gas Industry in Malaysia. The model generated offer insights into the technology process adoption scenario, and documents 12 factors arising which can prove useful in stimulating employees to go online and share.

Keywords: KMS adoption, positive factors, grounded theory.

1 Introduction

Nowadays, organizations experience a rapid growth in IT, and they use IT to manage their knowledge. However, an IT system alone does not guarantee successful knowledge management. Adoption of the IT innovation depends upon the human factor as well.

Despite the benefits of tools & technology in managing knowledge, only some organizations are successful in achieving their goals. For some companies the tools and technology provide benefits, yet there is a lack of adoption [1][2]. In addition, how tools and technology facilitate knowledge sharing and CoPs is still questionable [1][2][3][4]. When use of tools and technology is low, some organizations choose to abandon the innovation [5]. In this case study, active users of online knowledge sharing and virtual CoPs are only at 15%. Therefore, literature demands for qualitative research in providing an understanding and exploring some positive factors that influence KMS adoption within the particular context of Malaysia.

2 Background on the Research Area

2.1 Theories Used in IT Adoption

Fichman conducted the first review of IT adoption studies. He examined 18 studies conducted between 1981-1991, which asked questions related to improving technology assessment, adoption and implementation. The most widely accepted theory for IT adoption was the innovation diffusion theory of Rogers. Strongest results were noted when researchers examined: “(1) individual adoption, and/or (2)

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 167–176, 2011.

independent use technologies that impose a comparatively small knowledge burden on would-be adopters." These were instances in which the assumptions of innovation diffusion theory held [6].

In the literature on IT adoption from 1992-2003, 11 theories are noted as described in Table 1 below. Some of the studies examined individual adoption of IT, and others examined organizational adoption of IT. In this study, the researchers conducted both individual and organizational analysis of IT adoption.

Table 1. Theories used in individual and organizational IT adoption research (adopted from Jeyaraj, 2006)

Theory	Main Author(s)	Used in Individual Adoption Studies	Used in Organizational Adoption Studies
Theory of Reasoned Action	Fishbein & Ajzen (1975)	x	
Innovation Diffusion Theory	Rogers (1983, 1995)	x	x
Social Cognitive Theory	Bandura (1986)	x	
Diffusion/Implementation Model	Kwon & Zmud (1987)		x
Technology Acceptance Model	Davis (1989)	x	
Perceived Characteristics of Innovation	Moor & Benbasat (1991)	x	
Theory of Planned Behavior	Ajzen (1991)	x	
Tri-Core Model	Swanson (1994)		x
Technology Acceptance Model II	Venkatesh et al. (2003)	x	
Unified Theory of Acceptance and Use of Technology	Venkatesh et al. (2003)	x	
Critical Social Theory	Green (2005)	x	

To date IT adoption studies have examined the 135 factors, as seen in Appendix A [7].

2.1.1 Innovation Diffusion Theory

According to Rogers, an innovation is defined as, "An idea, practice, or object that is perceived as new by an individual or other unit of adoption" [8]. Hall and Khan described adoption as, "The choice to acquire and use a new invention or innovation." [9]. Rogers described diffusion as, "The process by which an innovation is communicated through certain channels over time among the members of a social system" [8].

Rogers proposed four main elements that influence the spread of a new idea: the innovation itself, communication channels, time, and a social system. A communication channel is the medium by which an individual communicates the message to another. Time is the period of making an innovation-decision. The rate of adoption is the speed at which members of a social system adopt the innovation. A social system is a set of interrelated units that are engaged in joint-problem solving for a common goal.

There are five steps in the process of innovation diffusion at an individual level. The original process of innovation diffusion identified: awareness, interest, evaluation, trial and adoption. The current process of innovation diffusion notes: knowledge, persuasion, decision, implementation and confirmation.

Awareness or knowledge is the stage at which an individual is exposed to an innovation. However, the person lacks information about the innovation and is not inspired to find more information about it. The interest or persuasion stage is when an individual is interested to know more information about an innovation. The evaluation or decision stage is when an individual will look at the advantages and disadvantages of the innovation. Then they will decide to adopt or to reject it. The trial and implementation stage is when an individual realizes the usefulness of an innovation and starts to look for more information about it. Employment of the innovation will vary depending upon the situation. The adoption and confirmation stage is when an individual decides to continue using the innovation and may use the innovation to its fullest potential.

Rogers suggested two factors that individuals consider when making a decision about an innovation: 1) Is the decision to adopt made freely and implemented voluntarily? 2) Who makes the decision? Based on these two factors, there are three types of innovation decisions: optional innovation-decisions, collective innovation-decisions and authority innovation-decisions. Optional innovation-decisions are made by those who are distinguished from others in a social system. Collective innovation-decisions are made by a group within a social system. Authority innovation-decisions are made by members of a social system, who are influenced by powerful others.

It is to be noted that there are several intrinsic characteristics of innovations that influence an individuals' decision to adopt or reject an innovation. Relative advantage means the improvement of the innovation over previous generations. Compatibility is when an innovation fits an individuals' life. Complexity is the ease or difficulty of use. Trialability is how easy and/or difficult it is for an innovation to be adopted. Observability is the visibility of the innovation to others.

As part of the discussion in the context of IT adoption studies the innovation diffusion theory is worthy of revisiting. The adoption S-shaped curve indicates the percentage of persons adopting an innovation on the y-axis and time on the x-axis, with the outcome being market saturation over time. This certainly appears to be the goal of all adoption-diffusion studies be they at the individual or the organizational level [8].

3 Methodology

Methodologically, this study employed GT. GT is a qualitative research method increasingly common in use in various disciplines. This method is recommended for hard sciences as well as social sciences [16]. Its application to information systems is very helpful for explaining phenomenon, developing context-based and process-oriented descriptions [10] [11] [12].

GT is a suitable approach for situations where researchers are trying to reveal participants' experiences, perceptions, and build a theoretical framework based on reality [13]. In this regard, the researchers would like to explore the employees' experiences and perceptions in real situations thus the data is revealed by the employees. As the research interest herein is to generate new insights for the existing literature and to understand in depth about the innovation adoption process of a KMS, the researchers employs an inductive approach of qualitative research by adapting the process and design of a GT approach instead of applying a deductive, hypothesis testing approach. This study is explorative and interpretive in nature. It looks into the concepts that build the innovation adoption process of a KMS. Therefore, a GT approach is most suitable to employ in this study for the following reasons.

The GT approach offers a set of procedures for coding and analyzing data, which keeps the analysis close to the data and presents the inductive discovery about the phenomena of the study. These procedures are structured and organized which leads the researchers to theory development [14]. As a result, the researchers are confident in the area of conceptualizing because it includes the resources of developing theoretical propositions from the data itself.

This study contributed to the research literature on GT by determining two new methodological process sequences as noted in Table 2 which innovatively combines the approaches of both Strauss and Glaser.

Table 2. Grounded theory methodology [15][16]

No.	GT Approach for This Study	Author
1	Start with having a general idea of where to begin.	(Strauss & Corbin, 2008)
2	Theoretical sensitivity comes from immersion in the data.	(Glaser, 1992)
3	Conceptual descriptions of situations.	(Strauss & Corbin, 2008)
4	The theory is grounded in the data.	(Glaser, 1992)
5	The credibility of the theory is from the rigor of the method.	(Strauss & Corbin, 2008)
6	The researcher is vigorous.	(Strauss & Corbin, 2008)
7	The data reveals the story.	(Glaser, 1992)
8	More rigorous coding and technique is defined. The nature of making comparisons diverges with the coding technique. Labels are carefully crafted at the time. Codes are derived from micro-analysis which analyzes data word by word.	(Strauss & Corbin, 2008)

The researchers conducted theoretical sampling and data collection process was constant and it is ceased when further data was no longer adding to the insights already gained. This indicator is called theoretical saturation. At this point, it was not necessary for further analysis because the analytical framework was saturated [15][17]. The further data of this study had not added new things therefore the theoretical model has been discovered at respondent number 8.

In terms of a process model of the analytic sequence of GT in this study (see Figure 1), the researchers explored in depth open, axial, and selective coding, and discovered conceptual process constructs of: bubbling, exploring, and arising.

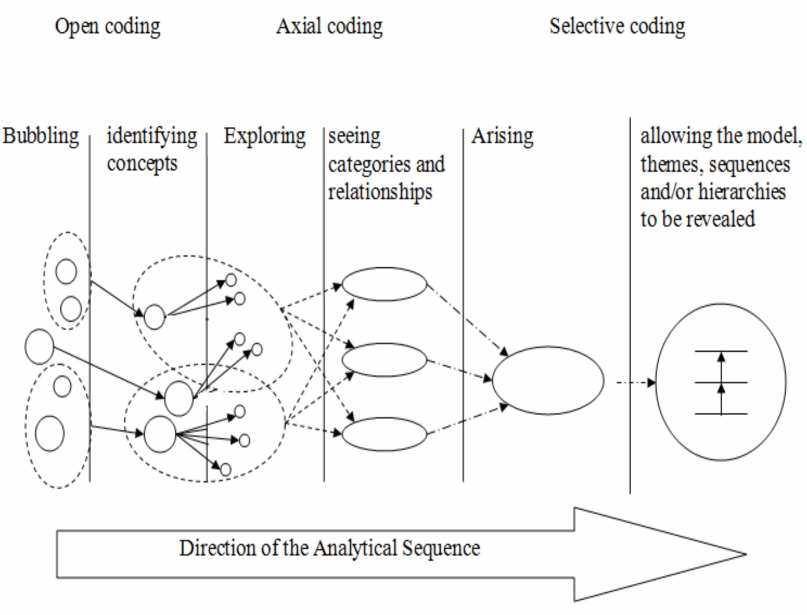

Open coding
Axial coding
Selective coding
Bubbling
identifying concepts
Exploring
seeing categories and relationships
Arising
allowing the model, themes, sequences and/or hierarchies to be revealed
Direction of the Analytical Sequence

Fig. 1. The GT analytical process in the data analysis (adapted from Warburton, 2005)[18]

Next, the researchers will describe and discuss the results of this study.

4 Results and Discussion

4.1 Demographic Findings

The demographic findings of this study are the participants' gender of 75% female and 25% male (see Figure 2). Participants job positions were 50% executives, 25% senior managers, and 25% managers (see Figure 3).

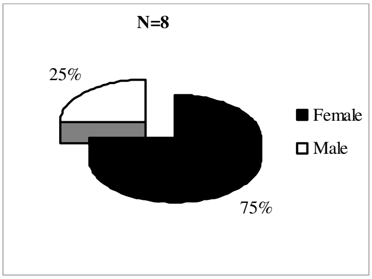


Fig. 2. Participants' Gender

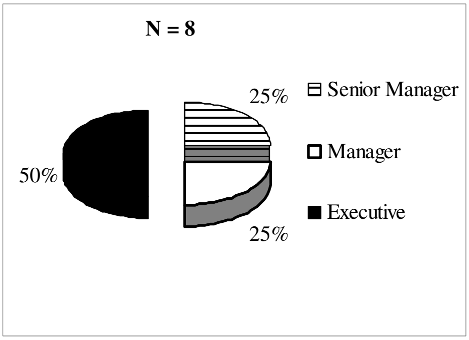


Fig. 3. Participants' Job Positions

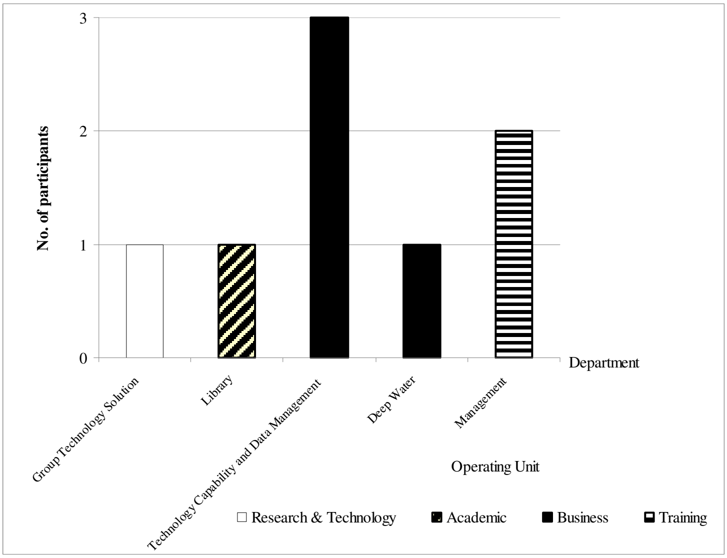


Fig. 4. Participants' Departments and Operating Units

The various management teams contributed very meaningful data to this study because of their knowledge and experiences with IT adoption, particularly the KMS. Figure 4 illustrates the distribution of the departments in which the participants worked and their operating units. The highest numbers of participants in this study were from the technology capability and data management department, which is under the business-operating unit.

4.2 Grounded Process of KMS Adoption

First and foremost this study derived the Grounded Process of KMS Adoption (see Figure 5). This model is a synthesis of the 8 models derived from the respondents and consists of three themes: Technology, Individual, and Process.

It is derived from and grounded in the data which explored the adoption of a KMS, at the leading oil and gas conglomerate of Malaysia with subsidiaries in 32 countries. At the Technological level, efficiency in terms of ease of use and being fast are important qualities, as well as, the technology satisfying a need and providing an experience of fulfillment. The experience of flow state proved essential at the Individual/People level along with self-benefit. The importance of knowledge sharing, organizational creativity and thus creation, and organizational growth are prominent

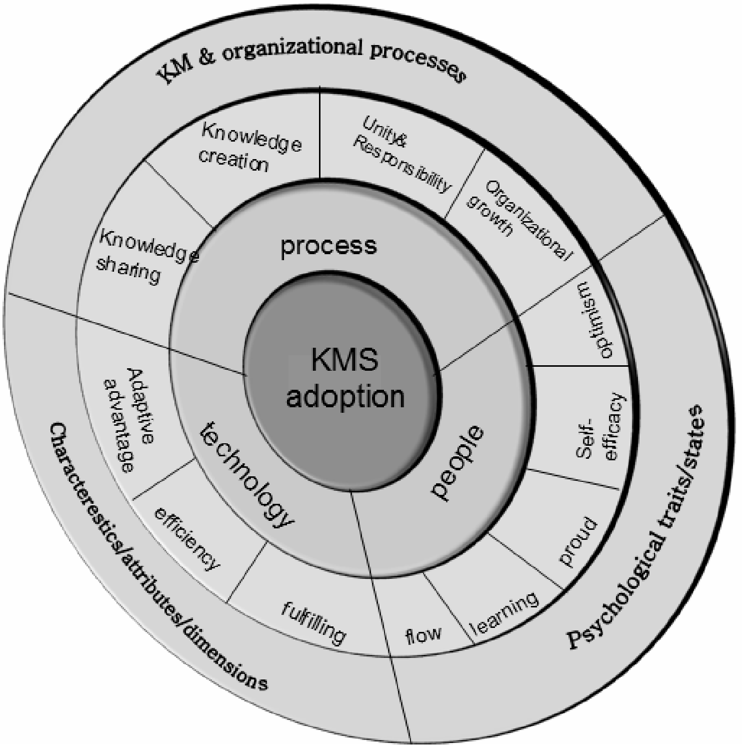


Fig. 5 Model of KMS Adoption: The Grounded Process

among the findings in the Process component of the model. Also, unique qualities of proud, optimism, unity and responsibility are noted which arise in part from the Islamic cultural values of the same nature, as the company in the case study is located in Malaysia, a primarily Islamic nation, with subsidiaries located in primarily Islamic nations around the world.

In addition, this study extends a new frontier by exploring the adoption of a KMS via the use of GT. Also, this study advanced the field of IT adoption studies by noting an additional 12 factors which arose from the data, see Table 3.

Table 3. Factors Findings

1. Adaptive Advantage
2. Arousal & Control
3. Creation
4. Efficiency
5. Fulfilling
6. Knowledge Sharing
7. Learning
8. Optimism
9. Organizational Growth
10. Proud
11. Responsibility & Unity
12. Self-benefit

5 Conclusion

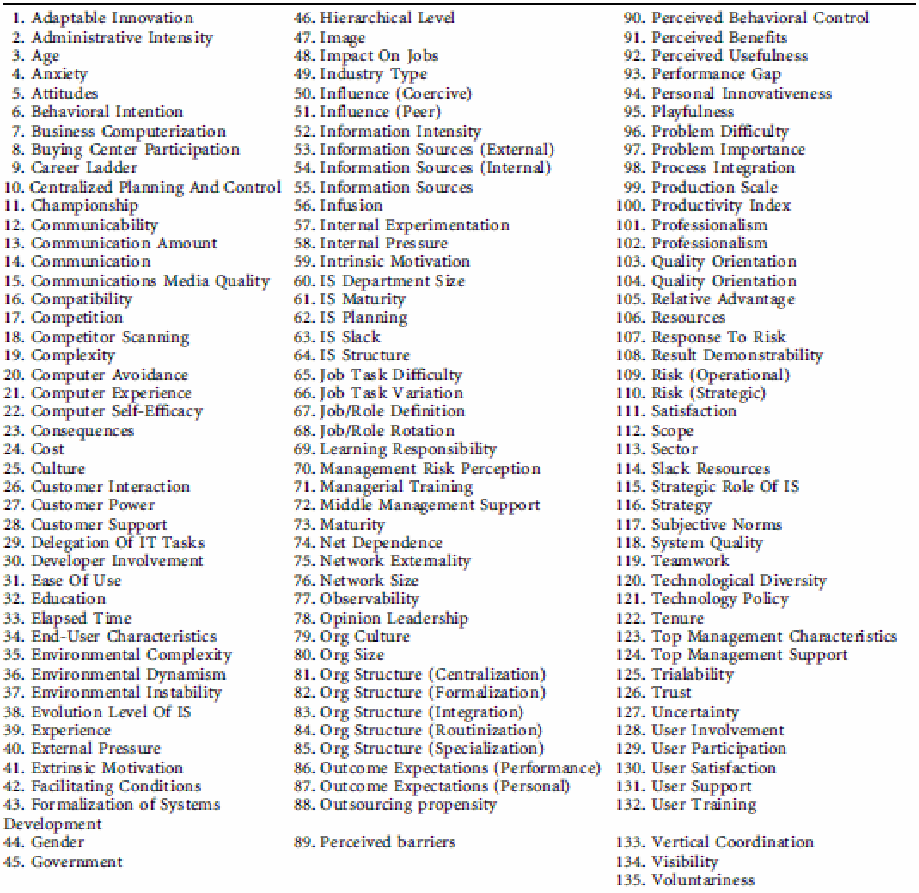
In conclusion, analysis of the adoption of KMS at company in case study revealed 15% active user at this time. Now, the company is facing the adoption gap and will have to call upon top management support to increase the rate of adoption. In order to stimulate employees to adopt system, as well as, to enhance knowledge in the field of IT adoption the researchers offers the following model for future scholars to explore (see Figure 5). In addition, this study advanced the field of IT adoption studies by noting an additional 12 factors, which arose from the data. The factors findings are: efficiency (ease of use/faster), fulfilling, adaptive advantage, flow (arousal and control), learning, proud, self-benefit, optimism, knowledge sharing, creation, organizational responsibility, organizational unity, and organizational growth.

References

1. Newell, S., Scarbrough, H., Swan, J., Hislop, D.: Intranets and Knowledge Management: Complex Processes and Ironic Outcomes. In: Proceedings of the 32nd Hawaii International Conference on System Sciences (1999)
2. Desouza, K.: Facilitating Tacit Knowledge Exchange. ACM 46(3) (June 2003)

3. Baalen, P., Ruwaard, J., Heck, E.: Knowledge Sharing in Emerging Networks of Practice: The Role of a Knowledge Portal. European Management Journal 23(3), 300–314 (2005)
4. Wang, C.Y., Yang, H.Y., Chou, S.T.: Using Peer-to-Peer Technology for Knowledge Sharing in Communities of Practices. Decision Support Systems 45(3), 528–540 (2008)
5. Stenmark, D.: Knowledge Sharing on a Corporate Intranet: Effects of Reinstating Web Authoring Capability. In: European Conference on Information Systems (ECIS) Proceedings (2005)
6. Fichman, R.: Information Technology Diffusion: A Review of Empirical Research. MIT Sloan School of Management 50 Memorial Drive, E53-314 Cambridge, MA 02139 (1992)
7. Jeyaraj, A., Rottman, J.W., Lacity, M.C.: A Review of the Predictors, Linkages, and Biases in IT Innovation Adoption Research. Journal of Information Technology 21, 1–23 (2006)
8. Rogers, E.M.: Diffusion of Innovations. Free Press, New York (1995)
9. Hall, B.H., Khan, B.: Adoption of New Technology. UCBERKELEY. Working Paper (May 2003)
10. Myers, M.: Qualitative Research in Information Systems. Management Information Systems Quarterly 21(2), 221–242 (1997)
11. Trauth, E.M.: In: Trauth, E. (ed.) Qualitative Research in Information Systems: Issues and Trends. Idea Group Publishing, Hershey (2001)
12. Urquhart, C.: An Encounter with Grounded Theory: Tackling the Practical and Philosophical Issues. In: Trauth, E. (ed.) Qualitative Research in Information Systems: Issues and Trends. Idea Group Publishing, London (2001)
13. Razavi, M., Iverson, L.: A Grounded Theory of Information Sharing Behavior in a Personal Learning Space. In: ACM CSCW, Banff, Alberta, Canada, November 4-6 (2006)
14. Charmaz, K.: Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. SAGE, India (2006)
15. Strauss, A.L., Corbin, J.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 3rd edn. Sage Publications, Inc., Los Angeles (2008)
16. Glaser, B.G.: Basics of Grounded Theory Analysis: Emergence Versus Forcing. Sociology Press (1992)
17. Glaser, B.: Basic Social Process. Grounded Theory Review 4, 1–27 (2005)
18. Warburton, W.I.: What are Grounded Theory Made of? In: Proceeding of the 2005 University of Southampton LASS Faculty Post-Graduate Research Conference, Southampton, UK, June 6-7 (2005)

Appendix A



1. Adaptable Innovation
2. Administrative Intensity
3. Age
4. Anxiety
5. Attitudes
6. Behavioral Intention
7. Business Computerization
8. Buying Center Participation
9. Career Ladder
10. Centralized Planning And Control
11. Championship
12. Communicability
13. Communication Amount
14. Communication
15. Communications Media Quality
16. Compatibility
17. Competition
18. Competitor Scanning
19. Complexity
20. Computer Avoidance
21. Computer Experience
22. Computer Self-Efficacy
23. Consequences
24. Cost
25. Culture
26. Customer Interaction
27. Customer Power
28. Customer Support
29. Delegation Of IT Tasks
30. Developer Involvement
31. Ease Of Use
32. Education
33. Elapsed Time
34. End-User Characteristics
35. Environmental Complexity
36. Environmental Dynamism
37. Environmental Instability
38. Evolution Level Of IS
39. Experience
40. External Pressure
41. Extrinsic Motivation
42. Facilitating Conditions
43. Formalization of Systems Development
44. Gender
45. Government
46. Hierarchical Level
47. Image
48. Impact On Jobs
49. Industry Type
50. Influence (Coercive)
51. Influence (Peer)
52. Information Intensity
53. Information Sources (External)
54. Information Sources (Internal)
55. Information Sources
56. Infusion
57. Internal Experimentation
58. Internal Pressure
59. Intrinsic Motivation
60. IS Department Size
61. IS Maturity
62. IS Planning
63. IS Slack
64. IS Structure
65. Job Task Difficulty
66. Job Task Variation
67. Job/Role Definition
68. Job/Role Rotation
69. Learning Responsibility
70. Management Risk Perception
71. Managerial Training
72. Middle Management Support
73. Maturity
74. Net Dependence
75. Network Externality
76. Network Size
77. Observability
78. Opinion Leadership
79. Org Culture
80. Org Size
81. Org Structure (Centralization)
82. Org Structure (Formalization)
83. Org Structure (Integration)
84. Org Structure (Routinization)
85. Org Structure (Specialization)
86. Outcome Expectations (Performance)
87. Outcome Expectations (Personal)
88. Outsourcing propensity
89. Perceived barriers
90. Perceived Behavioral Control
91. Perceived Benefits
92. Perceived Usefulness
93. Performance Gap
94. Personal Innovativeness
95. Playfulness
96. Problem Difficulty
97. Problem Importance
98. Process Integration
99. Production Scale
100. Productivity Index
101. Professionalism
102. Professionalism
103. Quality Orientation
104. Quality Orientation
105. Relative Advantage
106. Resources
107. Response To Risk
108. Result Demonstrability
109. Risk (Operational)
110. Risk (Strategic)
111. Satisfaction
112. Scope
113. Sector
114. Slack Resources
115. Strategic Role Of IS
116. Strategy
117. Subjective Norms
118. System Quality
119. Teamwork
120. Technological Diversity
121. Technology Policy
122. Tenure
123. Top Management Characteristics
124. Top Management Support
125. Trialability
126. Trust
127. Uncertainty
128. User Involvement
129. User Participation
130. User Satisfaction
131. User Support
132. User Training
133. Vertical Coordination
134. Visibility
135. Voluntariness

File Integrity Monitor Scheduling Based on File Security Level Classification

Zul Hilmi Abdullah[1], Nur Izura Udzir[1],
Ramlan Mahmod[1], and Khairulmizam Samsudin[2]

[1] Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia
43400 Serdang, Selangor
[2] Faculty of Engineering,
Universiti Putra Malaysia
43400 Serdang, Selangor

Abstract. Integrity of operating system components must be carefully handled in order to optimize the system security. Attackers always attempt to alter or modify these related components to achieve their goals. System files are common targets by the attackers. File integrity monitoring tools are widely used to detect any malicious modification to these critical files. Two methods, off-line and on-line file integrity monitoring have their own disadvantages. This paper proposes an enhancement to the scheduling algorithm of the current file integrity monitoring approach by combining the off-line and on-line monitoring approach with dynamic inspection scheduling by performing file classification technique. Files are divided based on their security level group and integrity monitoring schedule is defined based on related groups. The initial testing result shows that our system is effective in on-line detection of file modification.

Keywords: Operating System Security, Files Integrity, Monitoring Schedule, File Security Classification, Malicious Modification, HIDS.

1 Introduction

File integrity monitoring (FIM) is one of the security components that can be implemented in host environment. As a part of host based intrusion detection (HIDS) components, FIM should play a big role in detecting any malicious modification either from authorized or unauthorized users on their contents, access control, privileges, group and other properties. The main goal of related integrity checking or monitoring tools is to notify system administrators if any changed, deleted or added files detected [8]. File integrity checkers or monitors measure the current checksum or hash values of the monitored files with their original value.

In general, FIM can be divided into two categories, off-line and on-line monitoring scheme [7]. File system monitoring tools were originally used on their own before becoming a part of the intrusion detection system (IDS) when it is

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 177–189, 2011.

integrated with other components such as system logs monitoring, rootkits detection, and registry monitoring. System files as a core of the operating systems, contains information of the users, application, system configuration and authorization as well as program execution files [8]. Malicious modification of system file may cause disruption of the services or worse if it is used as a tool to attack other systems.

Recent solution of file integrity monitoring focusing on the on-line or real-time checking to enhance the capabilities of malicious modification detection. However, performance downgrade is a big issue in real time checking making it impractical for real world deployment. On the other side, higher cost of investment is required to deploy a new technology of integrity verification for the system such as hardware based protection mechanism using the Trusted Platform Module (TPM) which not only require TPM chips embedded on the computer hardware but also require an additional software to make it efficient.

The main target of the HIDS is to protect the operating system environment from intruders and unintended alteration or modification by authorized users. As one of the critical part in the operating system environment, the integrity of the system files must be put as high priority. However, to monitor all those system files in real-time is very difficult task and very costly especially for multi host and operating systems environment.

In this paper, we propose a software based file integrity monitoring by dynamically checking related files based on their sensitivity or security requirement. Sensitive files refer to the files which, if missing or improperly modified can cause unintended result to the system services and operation [23]. Classification of the sensitive and less sensitive files is used to determine the scheduling of the integrity monitoring of those files.

The rest of the paper is organized as follows: Section 2 discusses related works and compares our proposed techniques with these works. In Section 3, we describe our proposed system focusing on file security classification algorithm and FIM scheduling and how it differs with previous FIM. In Section 4, we quantify the initial implementation result of our algorithm in detecting file modification. This paper ended with discussion and conclusion in Section 5.

2 Related Work

In operating system environment, every component such as instruction, device drivers and other data is saved in files. There are huge number of files contained in modern operating system environment. Most of the time, files become a main target by the attackers to compromised the operating systems. The attack can be performed by modifying or altering the existence files, deletion, addition, and hide the related files. Many techniques can be implemented by the attackers to attack the files in the operating system environment and make file protection become a vital task. Implementation of FIM and other related system security tools is needed for that purpose.

As part of the HIDS functions, file integrity monitoring can be classified as off-line and on-line integrity monitoring. In the next section we discuss the off-line and on-line FIM followed by the multi platform FIM.

2.1 Off-Line File Integrity Monitoring

Tripwire [8] is a well known file integrity monitoring tool that motivates other researchers to develop more powerful FIM tools. Tripwire works based on four process, *init*, *check*, *update* and *test*. Comparing the current hash values of the files with the baseline values are the main principle of the FIM tools like Tripwire. However, relying on the baseline database require more maintenance cost due to more frequent system updates or patches [15]. In addition, off-line FIM needs to be scheduled in order to check the integrity of related files and most of the time can cause delay in detection of the modification. Samhain [19], AIDE [16], and Osiris [20] use the same approach too, so they also inherit almost the same issues as Tripwire.

Inspection frequency and the modification detection effectiveness is the main issue in the off-line FIM. In order to maintain the effectiveness of the FIM, high frequency inspection is needed at the cost of system performance, and vice versa. We overcome this issue by proposing a dynamic inspection scheduling by classifying related files to certain groups and the inspection frequency will vary between the group of files. Thus, from that approach, FIM can maintain its effectiveness with a more acceptable performance overhead to the system.

2.2 On-Line File Integrity Monitoring

On-line FIM is proposed to overcome the delay detection in off-line FIM approach by monitoring the security event involving system files in real-time. However, in order to work in real-time, it requires access of low level (kernel) activities which require kernel modification. When kernel modification is involved, the solution is kernel and platform-dependent, and therefore incompatible with other kernels and platforms.

As example, I3FS [12] proposed a real-time checking mechanism using system call interception and working in the kernel mode. However this work also requires some modification in protected machine's kernel. In addition, whole checksum monitoring in real time affected more performance degradation. I3FS offers a policy setup and update for customizing the frequency of integrity check. However it needs the system administrator to manually set up and update the file policy.

There are various on-line FIM and other security tools using the virtual machine introspection (VMI) technique to monitor and analyze a virtual machine state from the hypervisor level [13]. VMI was first introduced in Livewire [4] and then applied by the other tools like intrusion detection in HyperSpector [10] and malware analysis in Ether [3].

On the other side, virtualization based file integrity tools (FIT) has been proposed by XenFIT [15] to overcome the privileged issue on the previous user mode

FIT. XenFIT works by intercepting system call in monitored virtual machine (MVM) and sent to the privileged virtual machine (PVM). However, XenFIT requires a hardware virtualization support and only can fit with the Xen virtual machine, not other virtualization software. Another Xen based FIT is XenRIM [14] which does not require a baseline database. NOPFIT [9] also utilized the virtualization technology for their FIT using undefined opcode exception as a new debugging technique. However, all those real-time FIT only works on the Linux based OS.

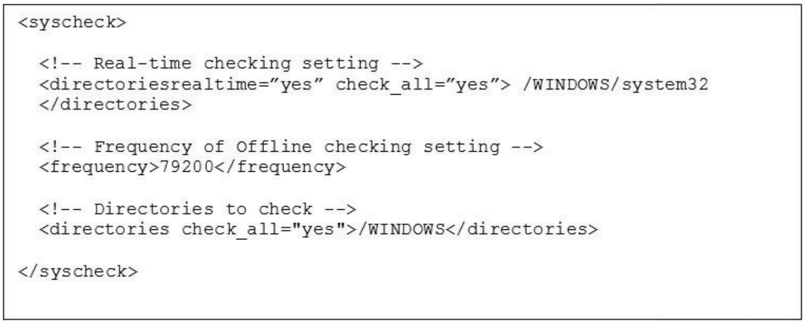
Another on-line FIM, VRFPS uses *blktap* library in Xen for their real time file protection tool [22]. This tool is also platform-dependent which only can be implemented in a Xen hypervisor. An interesting part in this tool is their file categorization approach to define which file requires protection and vice versa. We try to enhance their idea by doing the file classification to determine the scheduling process of file monitoring. VRFPS work on Linux environment in real time implementation but we implement our algorithm in Windows environment by combining on-line and off-line integrity monitoring. Combining the on-line and off-line integrity monitoring is to maintain the effectiveness of the FIM and to reduce the performance overhead.

2.3 Multi Platform File Integrity Monitoring

Developments in information technology and telecommunications led to higher demand for on-line services in various fields of work. Those services requires related servers on various platforms to be securely managed to ensure their trustworthiness to their clients. Distributed and ubiquitous environment require simple tools that can manage security for multi platform servers including the file integrity checking. There are a number of HIDS proposed to cater this need.

Centralized management of the file integrity monitoring is the main concern of those tools, and we take it as the fundamental features for our system and we focus more on the checking scheduling concern on the multi platform host. As other security tools have also implemented centralized management of their tools, such as anti-malware [18] and firewalls [2], FIM as part of HIDS also needs that kind of approaches to ensure the ease of administration and maintenance. We hope our classification algorithm and scheduling technique can also be applied to the other related systems.

Another issue to the FIM like Tripwire is the implementation on the monitored system which can be easily compromised if the attackers gain the administrator privilege. Wurster et al. [21] proposed a framework to avoid root abuse of file system privileges by restricts the system control during the installing and removing the application. Restricting the control is to avoid the unintended modification to the other files that not related to the installed or removed application. Samhain [19], and OSSEC [1] comes with centralized management of the FIT component in their host based intrusion detection system which allow multiple monitored systems to be managed more effectively. Monitoring the integrity of files and registry keys by scanning the system periodically is a common practice of the OSSEC. However, the challenge is to ensure the modification of related files can



```
<syscheck>

  <!-- Real-time checking setting -->
  <directoriesrealtime=”yes” check_all=”yes”> /WINDOWS/system32
  </directories>

  <!-- Frequency of Offline checking setting -->
  <frequency>79200</frequency>

  <!-- Directories to check -->
  <directories check_all="yes">/WINDOWS</directories>

</syscheck>
```

Fig. 1. Example of system integrity check configuration

be detected as soon as the event occurs as fast detection can be vital to prevent further damage.

OSSEC has features to customize rules and frequency of file integrity checking as shown in Figure 1. However it needs manual intervention by the system administrator. This practice becomes impractical in distributed and multi platform environment as well as cloud computing due to the large number of servers that should be managed. Therefore we try to implement multi platform FIM on the virtualized environment by customizing the scanning schedule with our techniques. Allowing other functions work as normal, we focus the file integrity monitoring features to enhance the inspection capabilities by scheduling it based on related files security requirements on related monitored virtual machines.

3 Proposed System

We found that most of the on-line and off-line FIM offer a policy setting features for the system administrator to update their monitoring setting based on the current requirement. However it can be a daunting task to the system administrator to define the appropriate security level for their system files especially those involving large data center. Therefore, a proper and automated security level classification of the file, especially system files, is required to fulfill this needs.

In this paper, we propose a new checking scheduling technique that dynamically update the file integrity monitoring schedule based on the current system requirement. This can be achieved by collecting information of related files such as their read/write frequency, owners, group, access control and other related attributes that can weight their security level. For initial phase, we only focus on the files owner and permission in our security classification.

Inspired by various services offered by modern operating systems, and multi services environments such as email services, web services, internet banking and others, the criticality of the integrity protection of those system is very crucial.Whether they run on a specific physical machine or in virtual environment, the integrity of their operating system files must be put in high priority to ensure the user’s trust on their services.

Centralized security monitoring is required to ensure the attack detection is still effective even though the monitored host has already been compromised. Windows comes with their own security tools such as Windows File Protection (WPC), Windows Resource Protection (WRP) and many more. However most of the tools rely on the privileged access of the administrator. If an attacker gains the administrator privileges, all modifications to the system files or other resources will look like a legal operation. So here where the centralize security monitoring is needed, when the critical resources are modified, the security administrator will be alerted although it is modified by local host administrator.

Identifying the most critical file that are often targeted by attackers is a challenging task due to the various techniques that can be used to compromise the systems. Based on the observation that specific attack techniques can be implemented to specific types of operating system services, we try to enhance the file integrity monitoring schedule by looking at the file security level for the specific host. It may vary from the other host and it can result dissimilarity type of scheduling but it is more accurate and resource-friendly since it fits on the specific needs.

3.1 System Architecture

The architecture of our proposed system is shown in Figure 2. The shaded area depicts the components that we have implemented. We develop our model based on the multi platform HIDS.

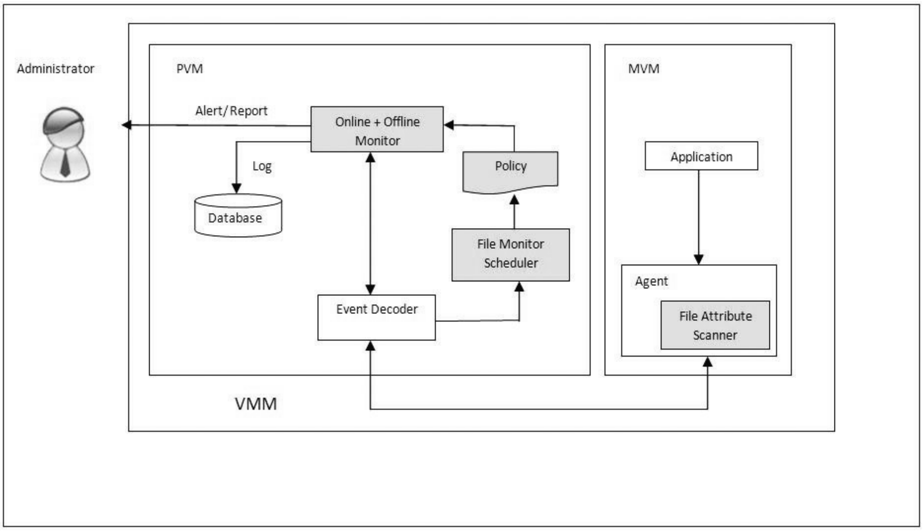


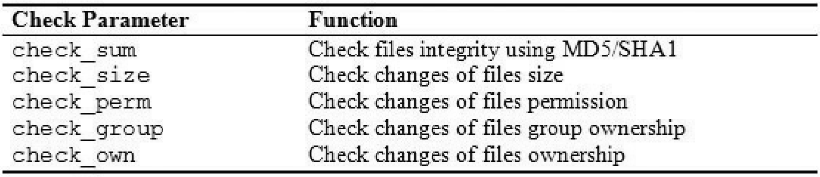
Fig. 2. Proposed FIM scheduling architecture

File Attribute Scanner (FAS). We collect file attributes to manipulate their information for our analysis and scheduler. Determining the specific group of files that require more frequent integrity inspection is a difficult task due to the various type of services offered by the operating systems. We assume that the system file structure is quite similar to various Windows based operating system. The security level of related group of files is the result of the combination between the file owner's rights and file permissions.

File attributes scanner (FAS) is locate in the agent packages that is deployed in MVM. In the FAS, files are scanned for the first time after our system installation on the MVM to create the baseline database. The baseline database of the files is stored in the PVM. In this process, the initial scheduler is created and added to the file monitor scheduler (FMS), which will overwrite the default policy. The monitoring engine will check the related files based on the defined policy. Then, if any changes occur in related files owner and permission, the FAS will update the classification and scheduler database.

We highlighted the FAS because it is what we have added in the previous agent's components. Another agent component is the file integrity monitor (FIM) that runs as the daemon process. FIM monitors the changes of the file content using the MD5 and SHA-1 checksum as well as changes in file ownership and permission. Event forwarding is part of the agent component which notifies the server for any event regarding file modification. Agent and server communicates via encrypted traffic.

Table 1. FIM check parameter



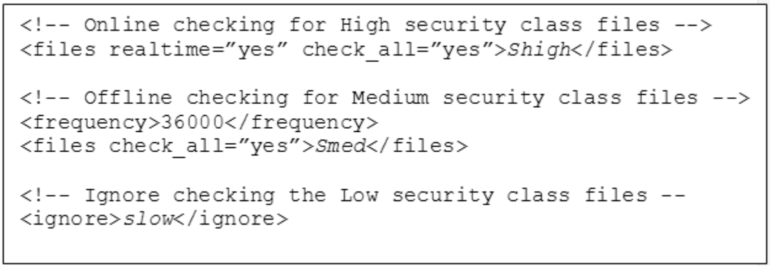
Check Parameter	Function
check_sum	Check files integrity using MD5/SHA1
check_size	Check changes of files size
check_perm	Check changes of files permission
check_group	Check changes of files group ownership
check_own	Check changes of files ownership

We implement our algorithm based on the OSSEC structure, hence, we also use the same check parameter suppose to (in Table 1) as OSSEC [6].

File Monitor Scheduler. File monitor scheduler (FMS) is one of our contributions in this paper. FMS collects file information from FAS in MVM via the event decoder to perform the file monitoring schedule based on the classification criteria. FMS has its own temporary database which contains groups of file names captured from FAS. The file groups will be updated if any changes occur in MVM captured by FAS. FMS will generate the FIM schedule and overwrite the default configuration file in the monitor engine. The monitoring engine will check related files based on the policy setting.

Policy. In default configuration, there are many built-in policy files which can be customized based on user requirements. In our case, we leave other policies as default configuration, but we add new policy enhancement on the FIM frequency. Our FIM policy relies on the file security level classification which is based on file ownership and permission captured on MVM. We offer dynamic policy updates based on our FMS result. The frequency of the policy update is very low due to infrequent changes in the file security level.

Monitoring Engine (On-line and Off-line Monitor). Monitoring engine plays a key function for our system. It communicates with the event decoder in order to obtain file information from MVM and pass instructions to the agent in MVM. File information is needed in the monitoring process either in real time or periodic checking based on the policy setting (Figure 3). The monitoring engine should send instructions to the agent in MVM when it needs current file information to compare with the baseline databases especially for the off-line monitoring process.



```
<!-- Online checking for High security class files -->
<files realtime=”yes” check_all=”yes”>Shigh</files>

<!-- Offline checking for Medium security class files -->
<frequency>36000</frequency>
<files check_all=”yes”>Smed</files>

<!-- Ignore checking the Low security class files --
<ignore>slow</ignore>
```

Fig. 3. Classification based FIM monitoring policy

3.2 File Classification Algorithm

In operating system environment, system files can be vulnerable to malicious modifications especially when attackers obtain administrator privileges. Therefore system file is the major concern in the FIM. However there are other files that should also be protected especially when related systems provide critical services to each other, such as web hosting, on-line banking, military related system, and medical related systems. It is quite subjective to define which files are more critical than others since every system provide different services.

In addition, huge number of files in the operating system environment is another challenge to the FIM in order to effectively monitor all those file without sacrificing the system performance. Hence, for that reason, we propose a file classification algorithm that can help FIM and other security tools to define the security requirements of related files.

Hai Jin et al [7] classified the files based on their security level weight as follows:

$$w_i = \alpha * f_i + \beta_i * d_i \ (\alpha + \beta = 1).$$

They represent the w_i as the weighted value for file i, f_i shows the file i access frequency, and they describe the significance of the directory containing the file i with d_i. They measure the files and directory weighted on the Linux environment where w_i represent the sensitivity level of the files. The variables, α and β, relate to the proportion of the frequency and the significance of the directory.

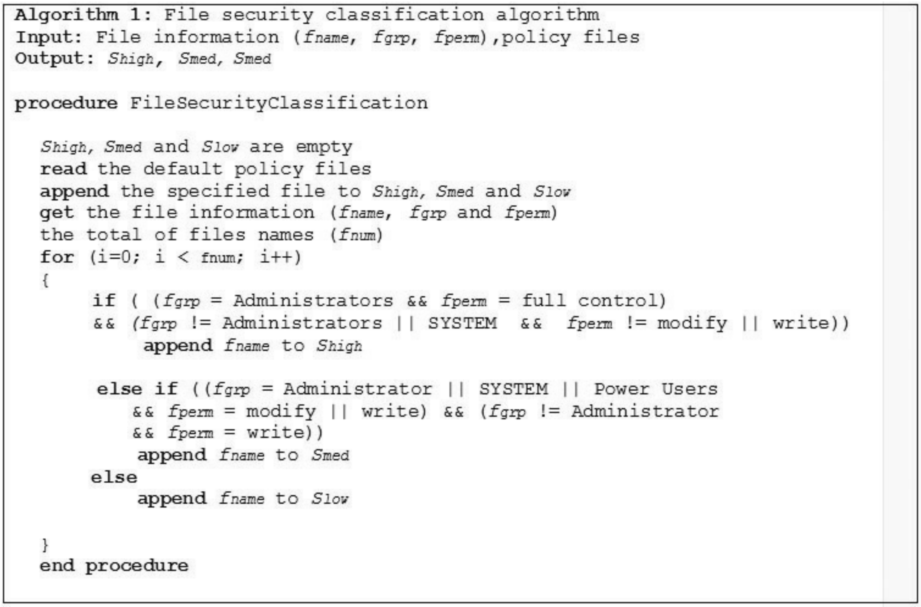
Microsoft offers File Classification Infrastructure (FCI) in their Windows Server 2008 R2 to assist users in managing their files [11]. FCI targets the business data files rather than system files. In other words, the files classification is based on the business impact and involves a more complex algorithm. Here we focus on the security impact on the systems and start with a simpler algorithm. In VRFPS file categorization, they classified the files the in Linux system into three types: *Read-only* files, *Log-on-write* files and *Write-free* files [22] to describe the security level of related files. In this paper, we also divide our file security level into three classes, *high*, *medium* and *low* security levels.

In this initial stage, we use the simple approach based on **user's right** and **object's permission** combination to define the file security level. However we exclude the user and group domains in this work as we are focusing more on the local files in MVM. User's rights refer to files owner that belong to a specific group that have specific privileges or action that they can or cannot perform. The files as objects that the user or group has permission or not to perform any operation to their content or properties [5]. For example, Ali as user, and a member of the Administrator group is permitted to modify the `system.ini` files contents. We define our files security level as follows:

- *High* security files: The files belong to Administrator. Other user groups have limited access to these files. Most of the system file type is in this group. This group of files requires on-line integrity checking.
- *Medium* security files: The files belong to Administrator group but other user groups also have permissions to read and write to these files. This group of file does not need on-line integrity monitoring but requires periodic monitoring, e.g. once a day.
- *Low* security files: The files are owned by users other than the Administrator group. This group of files can be ignored for integrity monitoring to reduce the system performance overhead during the monitoring process.

The goal of file security classification algorithm in Windows-based operating system is to dynamically schedule the integrity monitoring of those files. Different security levels of files need different monitoring schedules and this approach can optimize the FIM tool effectiveness and system performance as well. Moreover, the result of the file security classification provides information to the system administrator about the security needs of the related files.

Figure 4 shows our initial file security classification algorithm. We need basic file information including file names and its directory (*fname*), group of file's owner (*fgrp*), and file permission (*fperm*) as input, together with existing FIM policy files. All specified files will be classified as high (*Shigh*), medium (*Smed*) or low (*Slow*) security level based on their ownership and permission. Files' security



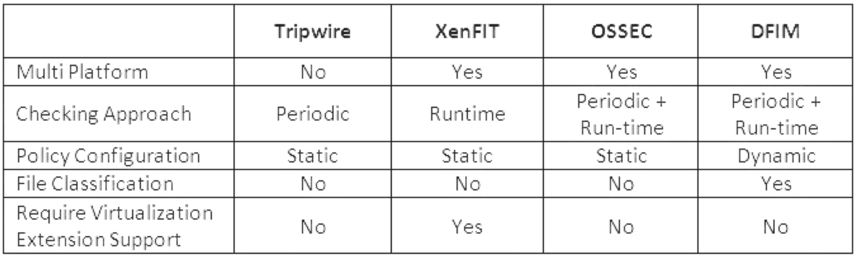
```
Algorithm 1: File security classification algorithm
Input: File information (fname, fgrp, fperm),policy files
Output: Shigh, Smed, Smed

procedure FileSecurityClassification

  Shigh, Smed and Slow are empty
  read the default policy files
  append the specified file to Shigh, Smed and Slow
  get the file information (fname, fgrp and fperm)
  the total of files names (fnum)
  for (i=0; i < fnum; i++)
  {
      if ( (fgrp = Administrators && fperm = full control)
      && (fgrp != Administrators || SYSTEM  &&  fperm != modify || write))
          append fname to Shigh

      else if ((fgrp = Administrator || SYSTEM || Power Users
         && fperm = modify || write) && (fgrp != Administrator
         && fperm = write))
          append fname to Smed
      else
          append fname to Slow

  }
  end procedure
```

Fig. 4. File security classification algorithm based on file ownership and permission

level information will be appended to the files information list, so any changes on their ownership and permission will be update. Dynamic update of the security level is needed due to discretionary access control (DAC) [17] implementation in Windows based OS which allow the file owner to determine and change access permission to user or group.

Table 2 indicate the comparison between our work with other FIM tools. We call our work as a dynamic file integrity monitoring (DFIM). The main objective of our work is to produce file integrity monitor in multi-platform environment. Variety of operating system in the needs more effective and flexible approaches. Therefore, base on some drawbacks of current FIM tools, we use file security classification algorithm to provide dynamic update of checking policy.

Table 2. Comparison with previous FIM tools



	Tripwire	XenFIT	OSSEC	DFIM
Multi Platform	No	Yes	Yes	Yes
Checking Approach	Periodic	Runtime	Periodic + Run-time	Periodic + Run-time
Policy Configuration	Static	Static	Static	Dynamic
File Classification	No	No	No	Yes
Require Virtualization Extension Support	No	Yes	No	No

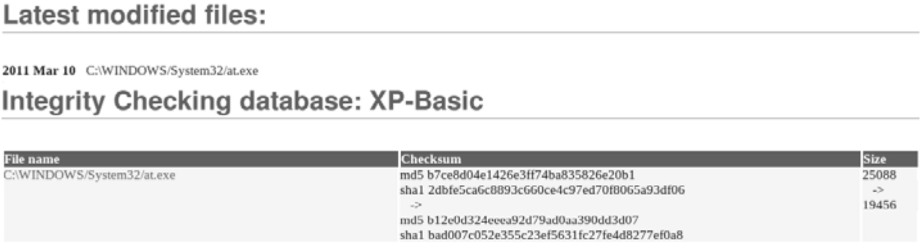
This is an initial work for file security classification in Windows environment and is not complete enough to secure the whole file in general. More comprehensive study will be carried out in future to enhance the file security classification algorithm for better result.

4 Experiment Environment

We tested our approach in the virtualized environment using Oracle Sun Virtualbox. Ubuntu 10 Server edition is installed as a management server or privileged virtual machine (PVM) for our FIM and Windows XP Service Pack 3 as a monitored virtual machine (MVM). We install HIDS for client server packages. The experiment environment is Intel Core2 Duo CPU E8400 with 3.0GHz, and 3GB memory.

We assume that the virtual machine monitor (VMM) provides strong isolation between PVM and MVM that fulfills the virtualization technology security requirement. Basically, our system does not require hardware-based virtualization support and it can be deployed on any CPU platform. However the algorithm can also be tested on other virtualization based FIT that relies on the hardware-based virtualization support such as XenFIT and XenRIM.

We tested our algorithm by doing some modification to the high security level files to measure the effectiveness of on-line FIM setting. We found that the modification can be detected immediately after the changes are made (Figure 5).



Latest modified files:

2011 Mar 10 C:\WINDOWS/System32/at.exe

Integrity Checking database: XP-Basic

File name	Checksum	Size
C:\WINDOWS/System32/at.exe	md5 b7ce8d04e1426e3ff74ba835826e20b1 sha1 2dbfe5ca6c8893c660ce4c97ed70f8065a93df06 -> md5 b12e0d324eeea92d79ad0aa390dd3d07 sha1 bad007c052e355c23ef5631fc27fe4d8277ef0a8	25088 -> 19456

Fig. 5. Detection of file modification

We are carrying out more detail experiments to measure the effectiveness of on-line and off-line FIM in detecting the file modification. In addition we will measure the performance overhead of our system to be compared to the native system.

5 Conclusion

We propose a new FIM scheduling algorithm based on file security classification that can dynamically update FIM needs. Most current FIM focus on their realtime FIM for sensitive files and ignored the other files without periodic checking their integrity. In addition, changes in file attributes are also ignored by most of FIM tools which can reduce their effectiveness. First, we try to simplify the different security groups for the files based on user's rights and object (file) permission combination. In Windows environment, DAC provides flexibility to the users to determine the permission setting of their resources. Changes to the object permission sometimes also require changes to their security requirement. Next, we will enhance the algorithm to develop more comprehensive classification of files security. Moreover, file security classification can be also used in other security tools to enhance their capabilities with acceptable performance overhead. Other platforms such as mobile and smart phone environments also can be a next focus in the file security classification in order to identify their security requirement. Lastly, centralized management of security tools should be implement due to the large number of systems owned by organizations to ensure security updates and patches can perform in a more manageable manner.

References

1. Ossec - open source host-based intrusion detection system, http://www.ossec.net/
2. Al-Shaer, E.S., Hamed, H.H.: Modeling and management of firewall policies. IEEE Transactions on Network and Service Management 1(1), 2 (2004)
3. Dinaburg, A., Royal, P., Sharif, M., Lee, W.: Ether: malware analysis via hardware virtualization extensions. In: CCS 2008: Proceedings of the 15th ACM Conference on Computer and Communications Security, pp. 51–62. ACM, New York (2008)
4. Garfinkel, T., Rosenblum, M.: A virtual machine introspection based architecture for intrusion detection. In: Proc. Network and Distributed Systems Security Symposium, pp. 191–206 (2003)
5. Glenn, W.: Windows 2003/2000/xp security architecture overview in expert reference series of white papers. Expert reference series of white papers, Global Knowledge Network, Inc. (2005)
6. Hay, A., Cid, D., Bary, R., Northcutt, S.: System integrity check and rootkit detection. In: OSSEC Host-Based Intrusion Detection Guide, Syngress, Burlington, pp. 149–174 (2008)
7. Jin, H., Xiang, G., Zou, D., Zhao, F., Li, M., Yu, C.: A guest-transparent file integrity monitoring method in virtualization environment. Comput. Math. Appl. 60(2), 256–266 (2010)
8. Kim, G.H., Spafford, E.H.: The design and implementation of tripwire: a file system integrity checker. In: CCS 1994: Proceedings of the 2nd ACM Conference on Computer and communications security, pp. 18–29. ACM, New York (1994)
9. Kim, J., Kim, I., Eom, Y.I.: Nopfit: File system integrity tool for virtual machine using multi-byte nop injection. In: Computational Science and its Applications, International Conference, vol. 0, pp. 335–338 (2010)
10. Kourai, K., Chiba, S.: Hyperspector: virtual distributed monitoring environments for secure intrusion detection. In: VEE 2005: Proceedings of the 1st ACM/USENIX International Conference on Virtual Execution Environments, pp. 197–207. ACM, New York (2005)
11. Microsoft. File classification infrastructure, technical white paper. Technical white paper (2009), http://www.microsoft.com/windowsserver2008/en/us/fci.aspx

12. Patil, S., Kashyap, A., Sivathanu, G., Zadok, E.: I3fs: An in-kernel integrity checker and intrusion detection file system. In: Proceedings of the 18th USENIX Conference on System Administration, pp. 67–78. USENIX Association, Berkeley (2004)
13. Pfoh, J., Schneider, C., Eckert, C.: A formal model for virtual machine introspection. In: VMSec 2009: Proceedings of the 1st ACM Workshop on Virtual Machine Security, pp. 1–10. ACM, New York (2009)
14. Quynh, N.A., Takefuji, Y.: A real-time integrity monitor for xen virtual machine. In: Proceedings of the International conference on Networking and Services, p. 90. IEEE Computer Society, Washington, DC, USA (2006)
15. Quynh, N.A., Takefuji, Y.: A novel approach for a file-system integrity monitor tool of xen virtual machine. In: ASIACCS 2007: Proceedings of the 2nd ACM Symposium on Information, Computer and Communications Security, pp. 194–202. ACM, New York (2007)
16. Rami, L., Marc, H., van den Berg Richard.: The aide manual, http://www.cs.tut.fi/~rammer/aide/manual.html
17. Russinovich, M.E., Solomon, D.A.: Microsoft Windows Internals. In: Microsoft Windows Server(TM) 2003, Windows XP, and Windows 2000 (Pro-Developer), 4th edn. Microsoft Press, Redmond (2004)
18. Szymczyk, M.: Detecting botnets in computer networks using multi-agent technology. In: Fourth International Conference on Dependability of Computer Systems, DepCos-RELCOMEX 2009, June 30- July 2, pp. 192–201 (2009)
19. Wichmann, R.: The samhain file integrity / host-based intrusion detection system (2006), http://www.la-samhna.de/samhain/
20. Wotring, B., Potter, B., Ranum, M., Wichmann, R.: Host Integrity Monitoring Using Osiris and Samhain. Syngress Publishing (2005)
21. Wurster, G., van Oorschot, P.C.: A control point for reducing root abuse of file-system privileges. In: CCS 2010: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 224–236. ACM, New York (2010)
22. Zhao, F., Jiang, Y., Xiang, G., Jin, H., Jiang, W.: Vrfps: A novel virtual machine-based real-time file protection system. In: ACIS International Conference on Software Engineering Research, Management and Applications, pp. 217–224 (2009)
23. Zhao, X., Borders, K., Prakash, A.: Towards protecting sensitive files in a compromised system. In: Proceedings of the Third IEEE International Security in Storage Workshop, pp. 21–28. IEEE Computer Society, Los Alamitos (2005)

A Convex Hull-Based Fuzzy Regression to Information Granules Problem – An Efficient Solution to Real-Time Data Analysis

Azizul Azhar Ramli[1,4], Junzo Watada[1], and Witold Pedrycz[2,3]

[1] Graduate School of Information, Production and Systems, Waseda University,
2-7, Hibikino, Wakamatsu, Kitakyushu, 808-0135 Japan
azizulazhar@moegi.waseda.jp, junzow@osb.att.ne.jp
[2] Department of Electrical and Computer Engineering, University of Alberta,
Edmonton, Alberta, Canada T6G 2V4
[3] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
pedrycz@ece.ualberta.ca
[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn
Malaysia, Parit Raja, 86400 Batu Pahat, Johor Darul Takzim, Malaysia

Abstract. Regression models are well known and widely used as one of the important categories of models in system modeling. In this paper, we extend the concept of fuzzy regression in order to handle real-time implementation of data analysis of information granules. An ultimate objective of this study is to develop a hybrid of a genetically-guided clustering algorithm called genetic algorithm-based Fuzzy C-Means (GA-FCM) and a convex hull-based regression approach being regarded as a potential solution to the formation of information granules. It is shown that a setting of Granular Computing helps us reduce the computing time, especially in case of real-time data analysis, as well as an overall computational complexity. We propose an efficient real-time information granules regression analysis based on the convex hull approach in which a Beneath-Beyond algorithm is employed to design sub convex hulls as well as a main convex hull structure. In the proposed design setting, we emphasize a pivotal role of the convex hull approach or more specifically the Beneath-Beyond algorithm, which becomes crucial in alleviating limitations of linear programming manifesting in system modeling.

Keywords: convex hull, Fuzzy C-Means, fuzzy regression, genetic algorithm, information granule.

1 Introductory Comments

Nowadays, we witness a significant growth of interest in Granular Computing (GrC) being regarded as a promising vehicle supporting the design, analysis and processing of information granules [1]. With regard of all processing faculties, information granules are collections of entities (elements), usually originating at the numeric level, that are arranged together due to their similarity, functional adjacency and indistinguishability or alike [1]. Given the similarity function to quantify the closeness between the

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 190–204, 2011.

samples, these data are clustered into certain granules, categories or classes [2]. The process of forming information granules is referred as information granulation.

GrC has begun to play important roles in bioinformatics, pattern recognition, security, high-performance computing and others in terms of efficiency, effectiveness, robustness as well as a structural representation of uncertainty [2]. Therefore, the need for sophisticated intelligent data analysis (IDA) tools becomes highly justifiable when dealing with this type of information.

Accordingly, the developed method discussed here exhibits good performance as far as computing time and an overall computation complexity are concerned. Fuzzy C-Means (FCM) clustering algorithm, introduced by Dunn in 1973 [3] and generalized by Bezdek in 1981, becomes one of the commonly used techniques of GrC when it comes to the formation of information granules [4] [5]. There has been a great deal of improvements and extensions of this clustering technique. One can refer here to the genetically-guided clustering algorithm called genetic algorithm-FCM (GA-FCM) and proposed by Hall *et al.* [3]. It has been shown that the GA-FCM algorithm can successfully alleviate the difficulties of choosing a suitable initialization of the FCM method. On the other hand, Ramli *et al.* proposed a real-time fuzzy regression model incorporating a convex hull method, specifically a Beneath-Beyond algorithm [6]. They have deployed a convex hull approach useful in the realization of data analysis in a dynamic data environment.

Associated with these two highlighted models (that is fuzzy regression and fuzzy clustering), the main objective of this study is to propose an enhancement of the fuzzy regression analysis for the purpose of analysis of information granules. From the IDA perspective, this research intends to augment the model given originally proposed by Bezdek by including the Ramli *et al.*'s approach. It will be shown that such a hybrid combination is capable of supporting real-time granular based fuzzy regression analysis.

In general, the proposed approach helps perform real-time fuzzy regression analysis realized in presence of information granules. The proposed approach comprises four main phases. First, we use the GA-FCM clustering algorithm to granulate the entire data set into a limited number of chunks –information granules. The second phase consists of constructing sub convex hull polygons for the already formed information granules. Therefore, the number of constructed convex hulls should be similar to the number of identified information granules. Next, main convex hull is constructed by considering all sub convex hulls. Moreover, the main convex hull will utilize the outside vertices which were selected from the constructed sub convex hulls. Finally, in the last phase, the selected vertices of the main constructed convex hull, which covers all sub convex hull (or identified information granules), are used to build a fuzzy regressions model. To illustrate the efficiency and effectiveness of the proposed method, a numeric example is presented.

This study is structured as follows. Section 2 serves as a concise and focused review of the fundamental principles of GrC as well as GA-FCM. Section 3discusses some essentials of fuzzy linear regression augmented by the convex hull approach. Section 4 discusses a processing flow of the proposed approach yielding real-time granular based fuzzy regression models. Section 5 is devoted to a numerical experiment. Finally, Section 6 presents some concluding remarks.

2 Introduction of Granular Information

Granular Computing (GrC) is a general computing paradigm that effectively deals with designing and processing information granules. The underlying formalism relies on a way in which information granules are represented; here we may consider set theory, fuzzy sets, rough sets, to name a few of the available alternatives [1]. In addition, GrC focuses on a paradigm for representing and processing information in a multiple level architecture. Furthermore, GrC can be viewed as a structured combination of algorithmic and non- algorithmic aspects of information processing [4].

Generally, granular computing is a twofold process: granulation and computation, where the former transforms the problem domain to one with granules, whereas the latter computes these granules to solve the problem. Granulation of information is an intuitively appealing concept and appears almost everywhere under different names, such as chucking, clustering, partitioning, division or decomposition [7]. Moreover, the process of granulation and the nature of information granules imply certain formalism that seems to be the most suited to capture the problem at hand. Therefore, to deal with the high computational cost which might be caused by a huge size of information granule patterns, we detailed here FCM algorithm which is one of commonly selected approaches to data clustering.

In general, the problem of clustering is that of finding a partition that captures the similarity among data objects by grouping them accordingly in the partition (or cluster). Data objects and functional within a group or cluster should be similar; data objects coming from different groups should be dissimilar. In this context, FCM arises as a way of formation of information granules represented by fuzzy sets [4]. Here, we briefly discuss the FCM clustering algorithm. The FCM algorithm minimizes the following objective function

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (U_{ik})^m D_{ik}^2(\mathbf{v}_i, \mathbf{x}_k) \tag{1}$$

where $U \in M_{fcn}$ is a fuzzy partition matrix, $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c)$ is a collection of cluster centers *(prototypes).* $\mathbf{v}_\mathrm{i} \in \Re^K \; \forall i$ and $D_{ik}(\mathbf{v}_i, \mathbf{v}_k)$ is a distance between x_k and the i^{th} prototype while m is a fuzzification coefficient, $m > 1$.

The FCM optimizes (1) by iteratively updating the prototypes and the partition matrix. More specifically, we choose some values of c, m and ε (termination condition – a small positive constant), then generate a random fuzzy partition matrix U^0 and set an iteration index to zero, $t = 0$. An iterative process is organized as follows. Given the membership value $\mu_{ik}^{(t)}$, the cluster centers $\mathbf{v}_i^{(t)} (i = 1, \ldots, c)$ are calculated by

$$\mathbf{v}_i^{(t)} = \frac{\sum_{k=1}^{n} (\mu_{ik}^{(t)})^m \mathbf{x}_k}{\sum_{k=1}^{n} (\mu_{ik}^{(t)})^m}. \tag{2}$$

Given the new cluster centers $\mathbf{v}_i^{(t)}$, we update the membership values of the partition matrix $\mu_{ik}^{(t)}$

$$\mu_{ik}^{t+1} = \left[\sum_{j=1}^{c} \left(\frac{\left\| \mathbf{x}_k - \mathbf{v}_i^{(t)} \right\|}{\left\| \mathbf{x}_k - \mathbf{v}_j^{(t)} \right\|} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{3}$$

This process terminates when $\left| U^{(t+1)} - U^{(t)} \right| \le \varepsilon$, or a predefined number of iterations have been reached [8]. In the following sub section, we present an enhancement of the FCM algorithm call GA-FCM, which used in this proposed research.

2.1 Genetically-Guided Clustering Algorithm (GA-FCM)

There are several studies employed genetic algorithm based clustering technique in order to solve various types of problems [9], [10], [11], [12]. More specifically, we exploit GA to determine the prototypes of the clusters in the Euclidean space $\Re^K$. At each generation, a new set of prototypes is created through the process of selecting individuals according to their level of fitness. In the sequel they are affected by running genetic operators [10], [12]. This process leads to the evolution of population of individuals that become more suitable given the corresponding values of the fitness function.

Basically, there are a number of research studies that have been completed which utilizing the advantages of GA-enhanced FCM. We focus here on the genetically guided clustering algorithm proposed by Hall *et al.*. Based on [3], in any generation, element i of the population is V_i, a $c \times s$ matrix of cluster centers (prototypes). The initial population of size P is constructed by a random assignment of real numbers to each of the s features of the c centers of the clusters. The initial values are constrained to be in the range (determined from the data set) of the feature to which they are assigned.

In addition, as V 's will be used within the GA, it is necessary to reformulate the objective function for FCM for optimization purposes. The expression (1) can be expressed in terms of distances from the prototypes (as done in the FCM method). Specifically, for $m > 1$ as long as $D_{jk}(v_j, x_k) > 0 \quad \forall j, k,$ we have

$$\mu_{ik} = 1 / \sum_{j=1}^{c} \left(\frac{D_{ik}(\mathbf{v}_i, \mathbf{x}_k)}{D_{jk}(\mathbf{v}_j, \mathbf{x}_k)} \right)^{2/(m-1)} \qquad \text{for} \quad 1 \le i \le c; 1 \le k \le n \tag{4}$$

This gives rise to the FCM functional reformulated as follows

$$R_m(V) = \sum_{k=1}^{n} \left(\sum_{i=1}^{c} D_{ik}^{1/(1-m)} \right)^{1-m} \tag{5}$$

to optimize R_m with a genetically-guided algorithm (GGA) [3]. Furthermore, there are a number of genetic operators, which relate to the GA-based clustering algorithm including *Selection*, *Crossover* and *Mutation* [3]. Moreover, the complete process of the genetically-guided algorithm (GGA) was discussed in [3].

3 A Convex Hull-Based Regression: A Review

In regression models, deviations between observed and estimated values are assumed to be due to random errors. Regression analysis is one of the common approaches used to describe relationships among the analyzed samples.

Regression explains dependencies between independent and dependent variables. The variables are called explanatory ones, when they are used to explain the other variable(s) [13] [14].

As an interesting and useful extension, Tanaka *et al.* introduced an enhancement of the generic regression model by generalizing the parameters of the regression model to be fuzzy numbers [15]. The models of this category of the model are reflective of the very nature of fuzzy relationships occurring between the dependent and independent variables. The model itself is expressed in the following form:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + \tilde{A}_2 X_2 + \ldots + \tilde{A}_K X_K \tag{6}$$

where $\mathbf{X} = [X_0, X_1, \ldots, X_K]$ is a vector of independent variables with $x_0 = 1$; $\tilde{\mathbf{A}} = [\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_K]$ is a vector of fuzzy coefficients represented in the form of symmetric triangular fuzzy numbers and denoted by $\tilde{A}_j = (\alpha_j, c_j)$ where α_j and c_j are the central value and the spread of the triangular fuzzy number, respectively.

From the computational perspective, the estimation of the membership functions of the fuzzy parameters of the regression is associated with a certain problem of linear programming (LP) [13].

Given the notation used above, (6) can be rewritten as

$$\tilde{Y}_i = (\alpha_0, c_0) + (\alpha_1, c_1) X_1 + (\alpha_2, c_2) X_2 + \ldots + (\alpha_K, c_K) X_K \tag{7}$$

where α_j and c_j $(j = 1,2,\ldots,K)$ are the center and the spread of the predicted interval of $\tilde{A}_j$, respectively.

The weaknesses of the implementation of the multi-dimensional fuzzy linear regression can be alleviated by incorporating the convex hull approach [6] [16].

In the introduced modification, the construction of vertices of the convex hull becomes realized in real-time by using related points (convex points) of the graph. Furthermore, Ramli *et al.* stated that the real-time implementation of the method has to deal with a large number of samples (data). Therefore, each particular analyzed sample stands for a convex point and is possibly selected as a convex hull vertex. Some edges connecting the vertices need to be re-constructed as well.

Let us recall that the main purpose of fuzzy linear regression is to form the upper and lower bounds of the linear regression model. Both the upper line Y^U and lower line Y^L for fuzzy linear regression are expressed in the form:

$$Y^U = \{A_0 + A_1 x_1 + \ldots + A_K x_K\}^U : \{\mathbf{A}\mathbf{x}_i^t\}^U = \boldsymbol{\alpha}\mathbf{x}_i^t + \mathbf{c}\left|\mathbf{x}_i^t\right| \tag{8}$$

$$Y^L = \{A_0 + A_1 x_1 + \ldots + A_K x_K\}^L : \{\mathbf{A}\mathbf{x}_i^t\}^L = \boldsymbol{\alpha}\mathbf{x}_i^t - \mathbf{c}\left|\mathbf{x}_i^t\right| \tag{9}$$

By using (8) and (9), we convert the problem to a general fuzzy regression that is similar to the one shown below:

i. Evaluation (objective) function.

$$\min_{\boldsymbol{\alpha},\mathbf{c}} \sum_{i=1}^{n} \sum_{j=2}^{K} c_j \left|P_{ij}\right|. \tag{10}$$

ii. Constraints

$$P_{i1} \in Y_i \Leftrightarrow \begin{cases} P_{i1} \le \alpha_0 + c_0 + \sum_{j=2}^{K} \alpha_j P_{ij} + \sum_{j=2}^{K} c_j \left|P_{ij}\right| \\ P_{i1} \ge \alpha_0 - c_0 + \sum_{j=2}^{K} \alpha_j P_{ij} - \sum_{j=2}^{K} c_j \left|P_{ij}\right| \\ \qquad (i = 1, \ldots, n) \end{cases} \tag{11}$$

The above expression can be further rewritten as follows:

$$\begin{aligned} Y^U &= \{Y_i^U \,|\, i = 1, \ldots, n\} \\ Y^L &= \{Y_i^L \,|\, i = 1, \ldots, n\}. \end{aligned} \tag{12}$$

We also arrive at the following simple relationships for P_{i1}

$$P_{i1} \le Y_i^U, \quad P_{i1} \ge Y_i^L \qquad (i = 1, \ldots., n) \tag{13}$$

It is well known that any discrete topology is a topology which is formed by a collection of subsets of a topological space X and the discrete metric ρ on X is defined as

$$\rho(x, y) = \begin{cases} 1 & \mathit{if}\ x \ne y \\ 0 & \mathit{if}\ x = y \end{cases} \tag{14}$$

for any $x, y \in X$. In this case, (X, ρ) is called a discrete metric space or a space of isolated points. According to the definition of discrete topology, expression (13) is rewritten as follows:

$$\begin{aligned} S(Y^U) &= \sum_{j=1}^{K} \{Y_j P_{ij}\}^U \geq 0 \\ S(Y^L) &= \sum_{j=1}^{K} \{Y_j P_{ij}\}^L \leq 0 \end{aligned} \tag{15}$$

where we assume that $P_{i1} = 1$.

This formula corresponds with the definition of the support hyperplane. Under the consideration of the range of

$$S \bigcap P \neq \phi \text{ and } P \subset S^+ \text{ or } P \subset S^-, \tag{16}$$

the following relation is valid:

$$\bigcap S(Y^U) = \bigcap S(Y^L). \tag{17}$$

This is explained by the fact that regression formula Y^U and Y^L are formed by vertices of a convex hull. Therefore, it is apparent that the constructed convex hull polygon or more specifically, its vertices clearly define the discussed constraints of fuzzy mathematical programming, becomes more reliable as well as significant for the subsequent processes.

Let us recall that the convex hull of a set S of points while $hull(S)$ is defined to be a minimum convex set containing S. A point $P \in S$ is an extreme point of S if $P \notin hull(S - P)$. Hence P denotes the set of points (input samples) and P_C is the set of vertices of the convex hull where $P_C \in P$. Therefore, the convex hull has to satisfy the following relationship:

$$conv(P) = conv(P_C) \tag{18}$$

Let us introduce the set

$$P_C = \left\{ x_{Cl} \in \Re^K \middle| l = 1, \ldots, m \right\} \subseteq P \tag{19}$$

where m is the number of vertices of the convex hull. Plugging this relationship into (11), we arrive at the following constraints:

$$P_{i1} \in Y_i \Leftrightarrow \begin{cases} P_{i1} \leq \alpha_0 + c_0 + \sum_{j=2}^{K} \alpha_j P_{ij} + \sum_{j=2}^{K} c_j \left| P_{ij} \right| \\ P_{i1} \geq \alpha_0 - c_0 + \sum_{j=2}^{K} \alpha_j P_{ij} - \sum_{j=2}^{K} c_j \left| P_{ij} \right| \\ \qquad (i = 1, \ldots, m). \end{cases} \tag{20}$$

By using (26), the constraints of the LP of the fuzzy linear regression can be written down in the following manner:

$$y_i \in Y_i \Leftrightarrow \begin{cases} y_i \leq \boldsymbol{\alpha}\mathbf{x}_i^t + \mathbf{c}\left|\mathbf{x}_i^t\right| \\ y_i \geq \boldsymbol{\alpha}\mathbf{x}_i^t - \mathbf{c}\left|\mathbf{x}_i^t\right| \\ (i = 1,\ldots,m). \end{cases} \tag{21}$$

Moreover, in order to form a suitable regression model based on the constructed convex hull, the connected vertex points are used as the constraints in the LP formulation of the fuzzy linear regression. Considering this process, the use of the limited number of selected vertices contributes to the minimized computing complexity associated with the model [6].

4 The Hybrid Combination of GA-FCM with Convex Hull-Based Regression Approach for Real-Time Regression

In general, there are four major components of this proposed approach. The description of related components is shows in table below, see Table 1. Furthermore, Fig. 1 below shows the synopsis of the entire processes.

Table 1. A Description of the main components of the proposed approach

No.	Component	Involved Algorithm/Processes	Description
1.	Genetically-Guided Clustering	GA-FCM algorithm	The used of GA-FCM algorithm for identify appropriate clusters which were represent information granules.
2.	Sub convex hull construction	Beneath-Beyond algorithm	Build a sub convex hull polygon for each identified cluster. This process will be repeated until all identified clusters achieved. The number of constructed convex hull should be same with constructed clusters.
3.	Convex hull construction	Beneath-Beyond algorithm	Build a convex hull polygon, which covers the whole constructed sub convex hull polygon.
4.	Fuzzy regression solution	LP formulation for fuzzy regression formulation	Used convex hull vertices in LP formulation of fuzzy regression for producing optimal models.

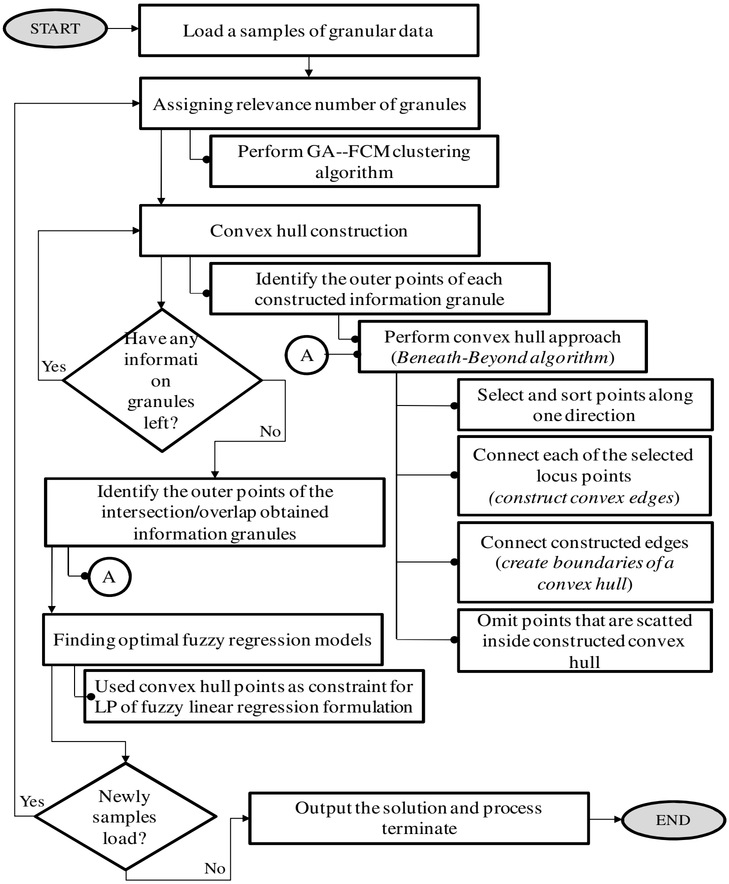

START
Load a samples of granular data
Assigning relevance number of granules
Perform GA--FCM clustering algorithm
Convex hull construction
Identify the outer points of each constructed information granule
Have any informati on granules left?
Yes
No
A
Perform convex hull approach (*Beneath-Beyond algorithm*)
Select and sort points along one direction
Connect each of the selected locus points (*construct convex edges*)
Connect constructed edges (*create boundaries of a convex hull*)
Omit points that are scatted inside constructed convex hull
Identify the outer points of the intersection/overlap obtained information granules
A
Finding optimal fuzzy regression models
Used convex hull points as constraint for LP of fuzzy linear regression formulation
Newly samples load?
Yes
No
Output the solution and process terminate
END

Fig. 1. A general flow of processing of the proposed approach

Referred on Fig. 2, we complete some iterations of the overall procedure considering that more data become available. Say, new samples are provided within a certain time interval, *e.g.*, they could be arriving every 10 seconds. Related to the comments made above, it becomes apparent that the quality of granular based fuzzy regression model can be improved by the hybrid combination of GA-FCM algorithm with convex hull-based fuzzy regression approach. The quality refers to the computing time as well as the overall computational complexity.

All in all, we do not have to consider the complete feature vectors for building regression models; just we utilize the selected vertices, which are used for the construction of the convex hull. As mentioned earlier, these selected vertices come from sub convex hull which represent appropriate information granules.

Therefore, this situation will lead to the decrease of computation load. On the other hand, related to the computational complexity factor for the subsequent iteration, it will only consider the newly added samples of data together with the selected vertices of the previous convex hull (main constructed convex hull polygon). For that reason, this

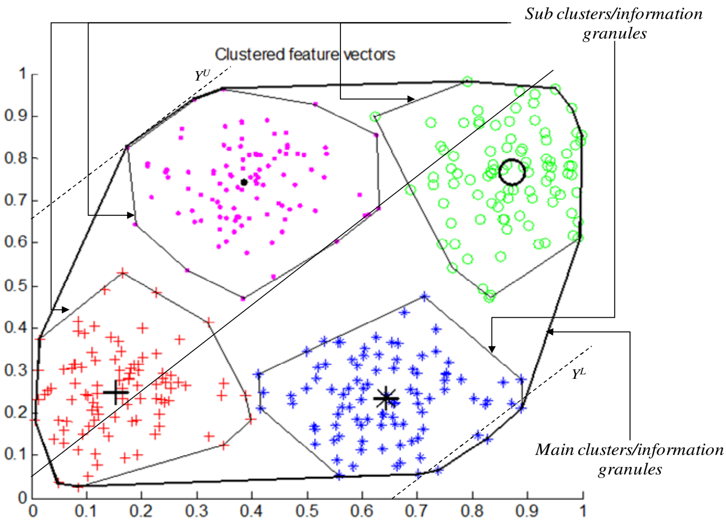


Fig. 2. An illustration of constructed sub-clusters and a main cluster

computing scenario will reduce the computational complexity because of the lower number of the feature vectors used in the subsequent processing of regression models.

5 A Numerical Example

We present a simple numerical example, which shows the efficiency of the proposed approach in the implementation of real-time granular based fuzzy regression. We assume that an initial group of samples consists of 100 data of the well-known *Iris* data set [17].

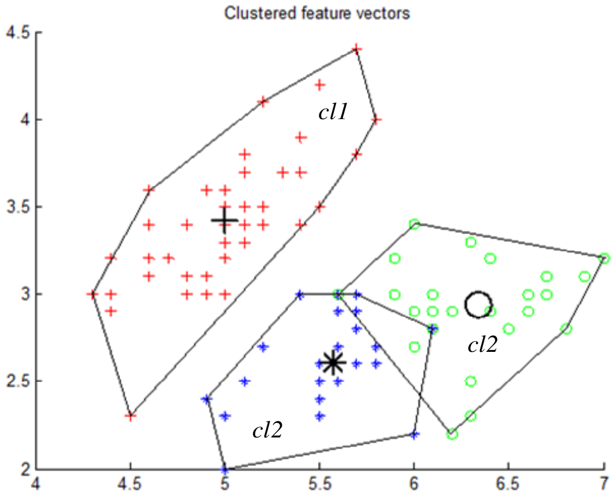


Fig. 3. Obtained clusters and constructed sub convex hulls for initial samples of data

Considering a distribution of these data, we construct sub convex hull polygons, which become the boundary of each identified cluster (or information granule), see Fig. 3.

Referring to the figure above, there are 3 constructed sub convex hulls. Table 2 covers the details of all clusters.

Table 2. Details of the obtained cluster along with the number of selected vertices for initial group of data samples

No.	Obtained Clusters	Selected Vertices
1.	Cluster 1 (*cl1*)	9
2.	Cluster 2 (*cl2*)	7
3.	Cluster 3 (*cl3*)	6

Next, we build a main convex hull which covers those sub convex hulls and among 22 of total selected clustered feature vectors (or loci points) as stated in Table 2, only 11 points were selected as convex hull vertices, see Fig. 4. In addition, these selected vertices are located as the outside points of the constructed clusters. By solving the associated LP problem that considered these selected vertices as a part of the constraint portion standing in the problem, we obtained the optimal regression coefficients, see below. We selected $h = 0.05$ to express goodness of fit or compatibility of data and the regression model

$$y = (2.071{,}0163) + (0.612{,}0.096)C1 + (0.639{,}0.075)C2 - (0.412, x0.000)C3$$

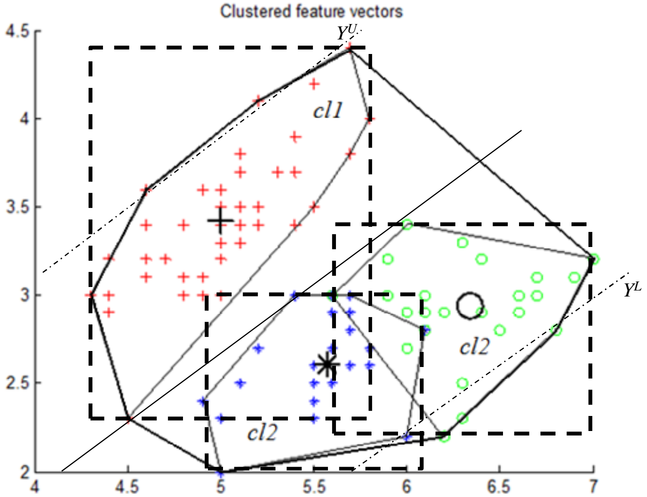


Fig. 4. Constructed of main convex hulls for initial samples of data

To deal with real-time scenario, we added a group of samples taken from the same data set, which consists of 50 patterns. In this case, we assume that an iteration process has been completed. Table 3 shows the details of each sub convex hull for initial group together with newly added data samples and Fig. 5 illustrate this related outcome.

Table 3. Detailed description of the clusters and the number of selected vertices for initial group together with newly added data samples

No.	Obtained Clusters	Selected Vertices
1.	Cluster 1 (*cl1*)	9
2.	Cluster 2 (*cl2*)	10
3.	Cluster 3 (*cl3*)	7

The total number of selected vertices for this newly data volume is 26 and out of them, the main constructed convex hull only used 10 vertices, Table 3. Finally, the obtained fuzzy regression model comes in the form.

$$y = (1.855, 0.173) + (0.651, 0.102)C1 + (0.709, 0.095)C2 - (0.556, 0.000)C3$$

while Fig. 6 shows the clustered feature vectors.

On the other hand, time-length recorded for initial samples of data (*first iteration*) is 00.28 seconds and for the second following iteration is only needs 00.09 seconds additional time-length. We can realize here, that although some number of samples added together with initial samples, the computational complexity as well as overall time consumption can be decreased.

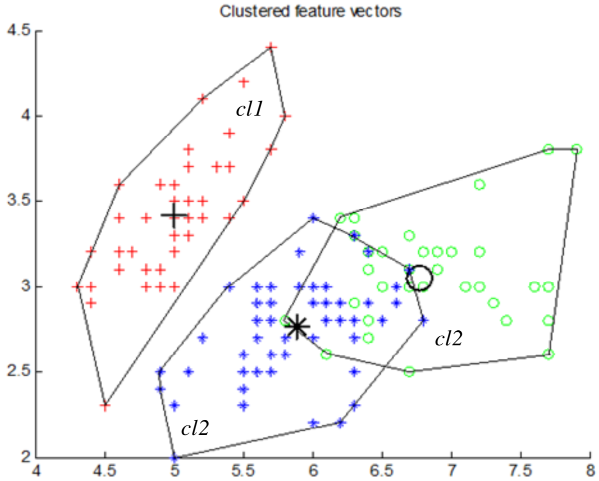


Fig. 5. Obtained clusters and constructed sub convex hulls for initial together with the newly added samples of data

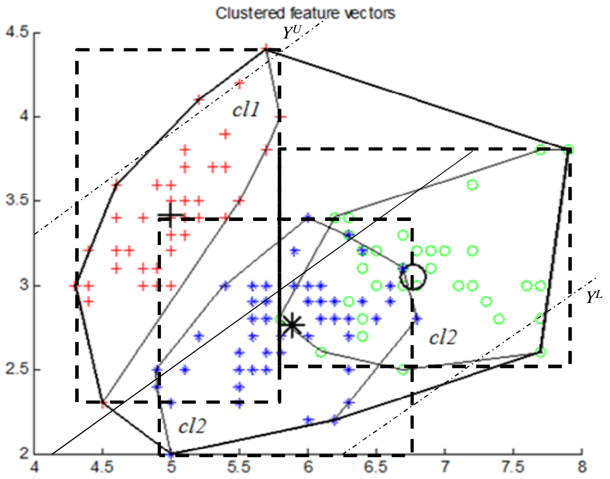


Fig. 6. Construction of main convex hulls for initial configuration together with newly added samples of data

6 Concluding Remarks

In this study, we have developed an enhancement of the IDA tool of fuzzy regression completed in the presence of information granules. Generally, the proposed approach first constructs a limited number of information granules and afterwards the resulting granules are processed by running the convex hull-based regression [6]. In this way, we have realized a new idea of real-time granular based fuzzy regression models being viewed as a modeling alternative to deal with real-world regression problems.

It is shown that information granules are formed as a result of running the genetic version of the FCM called GA-FCM algorithm [3]. Basically, there are two parts of related process, which utilize the convex hull approach or specifically Beneath-Beyond algorithm; constructing sub convex hull for each identified clusters (or information granules) and build a main convex hull polygon which covers all constructed sub convex hulls. In other word, the main convex hull is completed depending upon the outer plots of the constructed clusters (or information granules). Additionally, the sequential flow of processing was carried out to deal with dynamically increasing size of the data.

Based on the experimental developments, one could note that, this approach becomes a suitable design alternative especially when solving real-time fuzzy regression problems with information granules. It works efficiently for real-time data analysis given the reduced processing time as well as the associated computational complexity.

This proposed approach can be applied to real-time fuzzy regression problems in large-scale systems present in real-world scenario especially involving granular computing situation. In addition, each of the implemented phases, especially GA-FCM process and both sub and main convex hull construction processes have their own features in facing with dynamically changes of samples volume within a certain time interval. As a result, this enhancement (hybrid combination) provides an efficient platform for regression purposes, Although in this paper we dealt with small data sets (and this was done for illustrative purposes), it is worth noting that method scales up quite easily.

In further studies, we plan to expand the proposed approach by incorporating some other technologies of soft computing and swarm intelligence techniques.

Acknowledgments. The first author was supported by Ministry of Higher Education Malaysia (MOHE) under *Skim Latihan Akademik IPTA (SLAI-UTHM)* scholarship program at Graduate School of Information, Production and Systems (IPS), Waseda University, Fukuoka, Japan.

References

1. Bargiela, A., Pedrycz, W.: Granular Computing: An Introduction. Kluwer Academic Publishers, Dordrecht (2003)
2. Shifei, D., Li, X., Hong, Z., Liwen, Z.: Research And Progress of Cluster Algorithms Based on Granular Computing. International Journal of Digital Content Technology and its Applications 4(5), 96–104 (2010)
3. Hall, L.O., Ozyurt, I.B., Bezdek, J.C.: Clustering with a Genetically Optimized Approach. IEEE Transactions on Evolutionary Computation 3(2), 103–112 (1999)
4. Bargiela, A., Pedrycz, W.: Toward a Theory of Granular Computing for Humancentered Information Processing. IEEE Transactions on Fuzzy Systems 16(16), 320–330 (2008)
5. Chen, B., Tai, P.C., Harrison, R., Pan, Y.: FIK Model: Novel Efficient Granular Computing Model for Protein Sequence Motifs and Structure Information Discovery. In: Sixth IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2006), Arlington, Virginia, USA, pp. 20–26 (2006)
6. Ramli, A.A., Watada, J., Pedrycz, W.: Real-Time Fuzzy Regression Analysis: A Convex Hull Approach. European Journal of Operational Research 210(3), 606–617 (2011)
7. Höeppner, F., Klawonn, F.: Systems of Information Granules. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) Handbook of Granular Computing, John Wiley & Sons, Ltd., Chichester (2008), doi:10.1002/9780470724163.ch9
8. Nascimento, S., Mirkin, B., Moura-Pires, F.: A Fuzzy Clustering Model of Data and Fuzzy C-Means. In: IEEE Conference on Fuzzy Systems (FUZZ-IEEE2000), San Antonio, Texes, USA, pp. 302–307 (2000)
9. Alata M., Molhim M., Ramini A.: Optimizing of Fuzzy C-Means Clustering Algorithm using GA. World Academy of Science, Engineering and Technology, pp. 224–229 (2008)
10. Yabuuchi, Y., Watada, J.: Possibilistic Forecasting Model and Its Application to Analyze the Economy in Japan. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 151–158. Springer, Heidelberg (2004)
11. Lin, H.J., Yang, F.W., Kao, Y.T.: An Efficient GA-Based Clustering Technique. Tamkang Journal of Science and Engineering 8(2), 113–122 (2005)
12. Wang, Y.: Fuzzy Clustering Analysis by using Genetic Algorithm. ICIC Express Letters 2(4), 331–337 (2008)

13. Wang, H.-F., Tsaur, R.-C.: Insight of a Possibilistic Regression Model. Fuzzy Sets and Systems 112(3), 355–369 (2000)
14. Watada, J., Pedrycz, W.: A Possibilistic Regression Approach to Acquisition of Linguistic Rules. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) Handbook on Granular Commutation, ch. 32, pp. 719–740. John Wiley and Sons Ltd., Chichester (2008)
15. Tanaka, H., Uejima, S., Asai, K.: Linear Regression Analysis with Fuzzy Model. IEEE Transactions Systems, Man and Cybernetics 12(6), 903–907 (1982)
16. Ramli, A.A., Watada, J., Pedrycz, W.: Real-Time Fuzzy Switching Regression Analysis: A Convex Hull Approach. In: 11th International Conference on Information Integration and Web-based Applications and Services (iiWAS 2009), Kuala Lumpur, Malaysia, pp. 284–291 (2009)
17. Frank, A., Asuncion, A.: UCI Machine Learning Repository. School of Information and Computer Science. University of California, Irvine (2010), `http://archive.ics.uci.edu/ml`

Genetic Algorithm Approach to Path Planning for Intelligent Camera Control for Scientific Visualization

Djamalladine Mahamat Pierre and Nordin Zakaria

Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
nordinzakaria@petronas.com.my,
djamal2810@gmail.com
http://www.utp.edu.my

Abstract. In this paper, we propose to develop an intelligent camera control algorithm for scientific visualization. Intelligent camera control refers to a path planning algorithm that allows a virtual camera to navigate a scene autonomously. Intelligent camera overcomes some shortcomings of traditional manual navigation such as the risk of getting lost in the scene, or the user's distraction from the main goal of the study. In the past years, several path planning approaches have been proposed. While those approaches focus on determining the shortest path between two points, they cannot adapt to multiple constraints that a virtual camera is subjected to, in scientific visualization. Inspired by Unmanned Aerial Vehicle path planning, our algorithm uses genetic algorithm as an optimization tool. Finally, the paper presents the experimental results of our algorithm including an empirical study to determine the optimal values for the genetic parameters.

Keywords: Intelligent Camera Control, Genetic Algorithm, Path Planning.

1 Introduction

Scientific visualization is used in many areas ranging from chemistry to medicine, astronomy, fluid mechanics, and volume rendering. Many researches resulted in various approaches to improve scientific visualization. For intense, the term ”intelligent camera control [8]” is introduced by Steven M. Decker and refers to a virtual camera capable of exploring a virtual environment autonomously. The goal of the intelligent camera is to determine a collision free short pathway between the camera's current position and a destination within an acceptable time frame. In past years, various optimization approaches to path planning have been used to determine the shortest path between two points. The approaches include potential field, cubical path, Randomized Rapidly-Exploring Random Tree, neural network, and recently, genetic algorithm.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 205–213, 2011.

Steven M. Decker designed a framework for path finding using potential field [8]. While this approach does not clearly handle local minima, Decker, in his subsequent works in [9], proposed several ways to handle it. Although this approach efficiently provides the shortest path between two points, it is not robust and flexible to adapt to multiple constraints that the virtual camera can be subjected to, in visualization.

Potential field assigns attraction values to the objective (the destination point) while repulsive values are assigned to the obstacles. The camera is pulled by the attraction force at the objective, while it is pushed away from the obstacles by the repulsive forces. The CubicalPath, presented by Beckhaus et al. at [6] and [10] discretizes the scene into a cube space (coarse voxel) where attraction values assigned to the cubes dictate the motion of the camera. Both potential field and cubical path suffer from unhandled local minima and forces compensation problems. When those states occur, the camera is virtually stopped from reaching its goal.

Graham et. Al [4] exploited path finding algorithm using neural network. Leigh et. Al [5] describes an alternative to A*-Like path finding using genetic algorithm. Their algorithm appears more efficient with multiple agents. Unfortunately, it relies on 2D testing ground. Expending the use of the algorithm to 3D environment does not necessary produce the same success level.

Hugen et al.[11] provides a path planning approach for mobile robots using genetic algorithm. Sensors are used to locate the obstacles while the obstacle response is computed locally. The algorithm can fit in virtual environment with multiple moving obstacles. The only difference is the use of ray casting in virtual environment instead of sensors in real world. However, this algorithm has some shortcomings. While the robots operate in a 3D real world space, the path planning is computed for a 2D world assuming that the robots operate on a flat floor. Moreover, although the algorithm can operate in a completely unknown environment, it is not concerned with the visibility of the objective.

[1, 2, and 3] describe a path planning approach for Unmanned Aerial Vehicle (UAV). The approach relies on genetic algorithm for path optimization. Genetic Algorithm (HOLLAND, 1975) is a search heuristics that imitates evolution in nature. Although the algorithm presented in [1, 2, and 3] does not handle the early visibility of the objective, the virtual camera and the UAV have some similarities on their behaviours: finding a collision free pathway in a three-dimensional environment.

While most of the available paths planning algorithms focus on determining the shortest path, their applicability in scientific visualization is narrow. We are proposing an algorithm that takes into consideration, not only the length of the path, but also an early visibility of the objective of the camera, into consideration. In the following lines, objective refers to a point of interest in the virtual environment, which the camera should focus on at destination. The goal of our intelligent camera is to reach and set focus on the objective as early as possible.

2 Problem Statement

Manual navigation (using mouse, keyboard, joystick) is the most commonly used mode of navigation for virtual scene exploration. Although manual navigation gives user the freedom to explore the scene and learn, it has a lot of shortcomings:

- In a relatively complex scene, exploring the scene becomes difficult. Users may find it difficult to locate particular point of the scene (containing multiple obstacles).
- Manual navigation may distract users and take their attention away from the point of study.

Besides manual navigation, autonomous agents are also used to improve navigation within a virtual scene (intelligent camera control). However, those agents are more concerned about finding the shortest path rather than putting a particular interest on some areas of the virtual scene: which is one characteristic of scientific visualization. Figure 1 depicts two possible path ways for an intelligent camera. The upper path way is longer and allows the camera to view the objective at a time T_1. The lower pathway is shorter but allows the camera to view the objective only at time T_2 greater than T_1. This example shows that finding the shortest path for a virtual camera is not enough to fulfill the requirements for the scientific visualization.

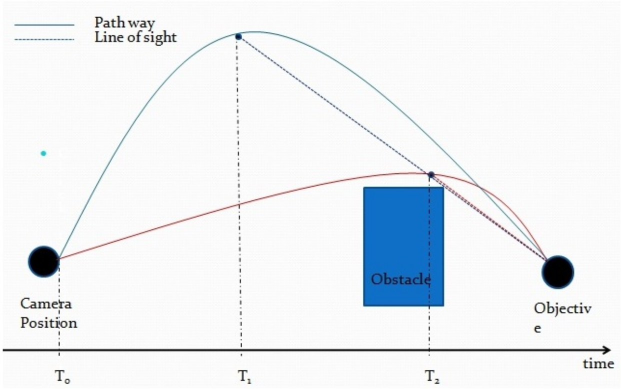


Fig. 1. Comparison of optimal path against shortest path

3 Objective

The objective of this research is to develop an algorithm capable of:

- Finding a collision free pathway from a virtual camera's current position to a destination.
- Ensure the pathway has an acceptable length and should be computed within an acceptable time frame.
- Keeping track of the destination point by ensuring that it is kept on the camera's field of view throughout the simulation.

4 Design and Implementation

4.1 Path planning

As a 'flying agent', our intelligent camera shares some similarities with a UAV. Therefore, the genetic algorithm based approach to path planning fits our algorithm better. The algorithm takes as parameters the camera's characteristics (position, orientation), the position of the objective, and the geometry of the scene (models and obstacles). The obstacles are created using triangle meshes. The path way is a B-Spline curve. The coordinates of the control points of the curve are passed to the GA as genes of a chromosome.

4.2 B-Spline

The phenotype of each individual is a B-Spline curve. A B-Spline curve is defined by a set of a few control points. The modification of the control points influence the shape of the curve. Those characteristics of the curve justify its choice over other curves'definitions which require more points, and therefore, require more memory space for computation. B-Spline curve is a parametric curve. The position of a point of the curve at a time t, in 3D space, is given by the following parametric function described [7]:

$$X(t) = \sum_{i=0}^{n} x_i B_{i,k}(t),\ Y(t) = \sum_{i=0}^{n} y_i B_{i,k}(t),\ Z(t) = \sum_{i=0}^{n} z_i B_{i,k}(t), \tag{1}$$

where $B_{i,k}$ is the blending function of the curve and k is the order. K represents the smoothness of the curve. A higher value of k provides a smoother curve. The curve representing the path of the camera is computed using a finite number, n+1, of control points . The value of t ranges from 0 to n-k-2 and is incremented at constant step. The definition of $B_{i,k}(t)$ is done recursively in term of knots value. Each knot is represented by the following function described in [1, 2, 9]:

$$Knot(i) = \begin{cases} 0, & \text{if } i < K \\ i - K + 1 & \text{if } K \le i \le n \\ n - K + 2 & \text{if } n < i \end{cases}$$

The definition of Bi, k (t) in term of the knot values is given by [1, 2, 9] as:

$$B_{i,1}(t) = \begin{cases} 1, & \text{if } Knot(i) \leq t < Knot(i+1) \\ 1, & \text{if } \begin{cases} Knot(i) \leq t \leq Knot(i+1) \\ \text{and} \\ t = n - K + 2 \end{cases} \\ 0, & \text{otherwise} \end{cases}$$

$$B_{i,k}(t) = \frac{(t - Knot(i)) \times B_{i,K-1}(t)}{(Knot(i+K-1) - Knot(i)} + \frac{(Knot(i+K) - t) \times B_{i+1,k-1}}{Knot(i+K) - Knot(i+1)}, \quad (2)$$

4.3 Genetic Algorithm

Genetic algorithm is a search heuristic that imitates evolution in nature. A potential solution, or individual, is represented by a chromosome. Each chromosome may contain several genes representing the characteristic of the individual. Starting from a couple of potential solutions, the individuals go through some transformations to generate a new and fitter individual. The transformations include reproduction, crossover and mutation. Reproduction involves making a copy of a chromosome. Crossover changes the content of individual by swapping the values of genes between two chromosomes. This approach mimics ‘mating’ of the individuals involved [12]. Mutation, on the other hand, alters the value of a gene to generate a new individual. If an individual is judged unfit in the process, it is simply discarded. The judgement on fitness of the individuals is based on a value (fitness value) computed using a fitness function. The fitness function is defined based on characteristics of the chromosome, or genes’ values.

While obstacles in [1, 2, and 3] are represented by a specific function, in our testing ground, they are represented by triangle meshes. Ray casting is used to determine the position of the obstacles and test the validity of the control points. The coordinates of the control points represent the genes of chromosomes or individuals. Chromosomes are evaluated using a fitness function f. The fitness function is inversely dependant on a sum of terms f_i. Each term f_i represents a penalty, or the extent to which the virtual camera is far from meeting a particular constraint. The following formula is a representation of the fitness f:

$$f = 1/\sum_{i=1}^{c} a_i f_i \quad (3)$$

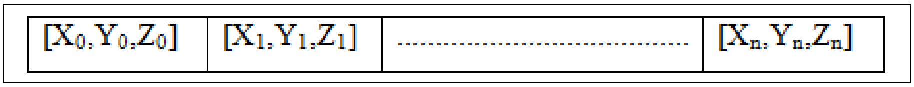
where c is the number of constraints, a_i is the weight of the term f_i. An individual with lower f_i values has higher fitness values; therefore, a camera following a pathway derived from such individual is closer to meet the constraints.

Two fixed points represent the camera's initial position and the objective respectively. Because of the static nature of the geometry of the scene and the objective, it is possible to determine one optimal path for the camera using a fixed number of control points. When the objective is static, four constraints apply to the optimization problem.

- f_1 penalizes all segments of the curve that cross the obstacles. The fitter curves are those with less penalties. f_1 is the most important constraint because it validates all the feasible points of the path. The penalty is computed using two successive points, P(t) and P(t+1) determined by (1). A ray, originated at P(t) is casted towards P(t+1). If there is no feedback, than there is no obstacle insight. If there is a feedback and the distance between the P(t) and the obstacle is greater than the distance between P(t) and P(t+1), no penalty is given. That is because the segment delimited by P(t) and P(t+1) is not crossing the triangle hit by the ray. On the other hand, if the distance is smaller, a penalty is given by incrementing the term f_1. The smaller distance indicates that the triangle of the obstacle hit by the ray is located between P(t) and P(t+1); in other words, the segment is crossing the obstacle.
- f_2 penalizes all points from which the objective is not visible. In other words, f_2 help getting a clear line of sight with the objective and strives to kept it. f_2 determines the early visibility of the objective.
- f_3 is a special penalty related to the length of the path way. A bigger value of f_3 reflects a longer path from the camera's position to the destination.
- f_4 ensures that a premature convergence of the control points does not occur. In order to avoid the accumulation of the control points in a location, a safe distance is maintained between two consecutive control points. f_4 indirectly penalizes local optima.

4.3.1 Genes

In our problem, the coordinates of the control points represent the genes. Since all the control points should be taken into consideration for the curve definition, the expected solution is a set of the same number of control points. If we have n+1 control points, then they are represented by genes indexed by 0 through n as depicted in the Figure 2.



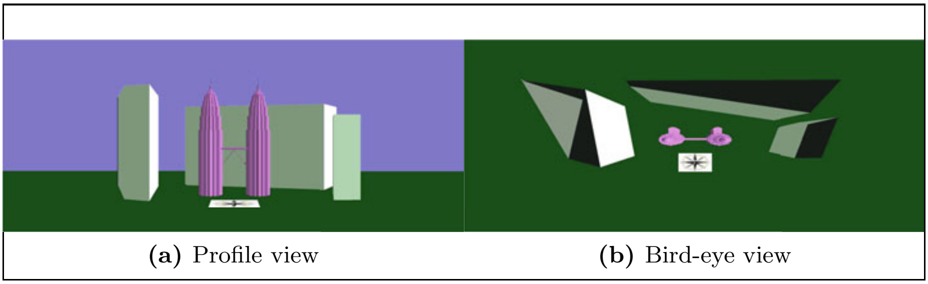
$[X_0,Y_0,Z_0]$	$[X_1,Y_1,Z_1]$	....	$[X_n,Y_n,Z_n]$

Fig. 2. Representation of a chromosome with a series of coordinates as its genes

X_i, Y_i, and Z_i, in Figure 2, represent the coordinates of the control points in 3D space coordinates.

5 Results

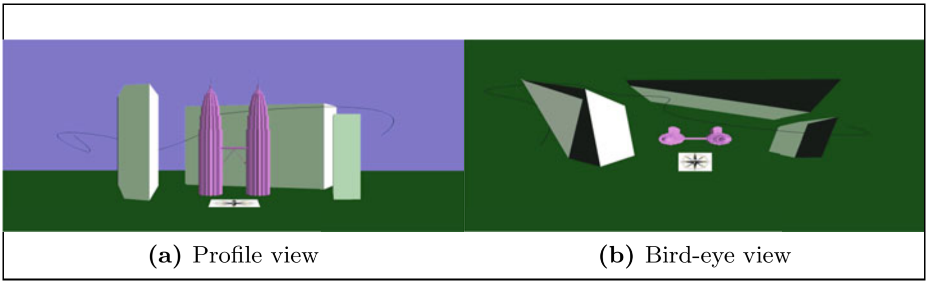
Our testing ground is a virtual environment with models. The models, representing the obstacles, are made of triangle meshes. Figure 3 shows a the virtual environment with buildings. The starting point and destination of the virtual camera are on either sides of the group of buildings.



(a) Profile view **(b)** Bird-eye view

Fig. 3. A two-perspective view of the virtual environment

With the default value of f_1 kept to 1, the maximum possible fitness that can be (logically) obtained is 1 (1/(1)). That is (logically) possible if and only if the penalties from all the terms f_2, f_3 and f_4 are null; in other words, there is no obstacle between the start point and the destination of the camera. A simple test is conducted to observe the phenotypic change of the pathway. With a crossover rate of 50 percent, a mutation rate of 20 percent, and the population size of ten (10), the fitness value of the fittest individual,after one hundred (100) generations, is 0.3264.



(a) Profile view **(b)** Bird-eye view

Fig. 4. Profile and Bird-eye views of a pathway with a fitness value of 0.3264

The relatively low fitness value can be justified by the length of the generated pathway as well as the obstraction of the objective from each discrete point of the curve. We can observe that the pathway links the start point to the destination without touching or crossing the models (buildings). However, The pathway sometimes presents unessary curves, which increases its lengths, and consequently, decreases its fitness. On the otherhand, from most discrete points of the pathway, it is impossible for the camera to perceive the objective, which also has a negative effect on the fitness.

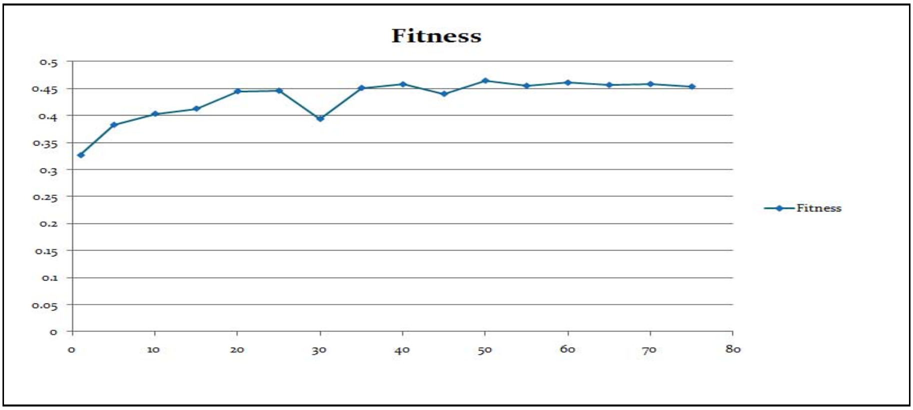


Fig. 5. Variation of fitness value with respect to the population size

6 Conclusion and Observation

There has been a lot of debate on what is the appropriate values for the crossover or mutation rates to have an optimal individual. Some experts suggest that, because those values may vary according to the problem in hand, tuning can be used to determine them. However, empirical studies conducted in our test case are inconclusive. First, the probabilistic characteristic of genetic algorithm and the random generation of genes during mutation or the creation of the population makes it hard to anticipate the impact of those values for the next test case. Dumitrescu et. al [13] states that a higher propulation size provides larger variety, and consequently, a bigger chance of exploiting the different individuals to determine a more reliable result. That statement is verified only when the individuals of the initial population are scattered and occupy a large space within the search space. But, if the initial population is generated randomly, there is guarantee that the individuals will be scattered within the search space. A series of tests, conducted to depict a relationship between the population size and the fitness values, is shown at Figure 2.

A general view of the graph shows that the fitness value increases as the size of the population rises. However, the sudden drops of the fitness at 30, 45 and 55 proves the assumption wrong. This issue might be solved if the population is systematically scattered in the search space. In order to maintain consistancy, it is imperative to consider modifying the selection procedure. In a traditional genetic algorithm, the individuals that are subjected to genetic transformations are selected randomly. This may give the perception that the final result is not predictable. The main question remains if the radom characteristics of the selection procedure and the population creation is necessary. In a typical video games, it is important that the agents' behaviors are unpredictable. However, the same cannot be said about path planning as users would not mind if the camera uses the same pathway at different test cases as long as the fitness of the pathway is acceptable.

References

1. Nikolos, I.K., Tsourveloudis, N.C., Valavanis, K.P.: Evolutionary Alogrithm Based O_-Line Path Planner for UAV. AUTOMATIKA 42(3-4), 143–150 (2001)
2. Nikolos, I.K., Valavanis, K.P., Member, S., Tsourveloudis, N.C., Kostaras, A.N.: Evolutionary Alogrithm Based Of- ine/Online Path Planner for UAV Navigation. IEEE Transsactions on Systems, Man, nad Cybernetics-Part B. 33(6) (2003); Euro-Par (2006)
3. Nikolos, I.K., Tsourveloud, N.C., Valanis, K.P.: EvolutionaryAlgorithm Based 3-D Path Planner for UAV Navigation
4. Graham, R., McCabe, H., Sheridan, S.: Neural Networks for Real-time Path finding in Computer Games. School of Informatics and Engineering, Institute of Technology at lanchardstwon, Dublin 15
5. Leigh, R., Louis, S.J., Miles, C.: Using a Genetic Algorithm to Ex- plore A*-like Pathfinding Algorithms. In: Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games, CIG 2007 (2007)
6. Beckhaus, S., Ritter, F., Strothotte, T.: CubicalPath - Dynamic Potential Fields for Guided Exploration in Virtual Environments
7. Mittal, S., Deb, K.: Three-Dimensional Offline Path Planning for UAVs Using Multiobjective Evolutionary Algorithms. In: IEEE Congress on Evolutionary Computation, CEC 2007 (2007)
8. Drucker, S.M.: Intelligent Camera Control for Graphical Environments. In: partial fulfillment of the requirements for the degree of Doctor of Philosophy at Massachusetts Institue of Technology (June 1994)
9. Drucker, S.M., Zeltzer, D.: Intelligent Camera Control in a Virtual Environment
10. Beckhaus, S.: Dynamic Potential Fields for Guided Exploration in Virtual Environments. Dissertation (2002)
11. Burchardt, H., Salomon, R.: Implementation of Path Planning using Genetic Algorithms on Mobile Robots
12. Rotsan, N., Meffert, K.: JGAP Frequently asked questions. Copyright (2002- 2007)
13. Dumitrescu, D., Lazzerini, B., Jain, L.C., Dumitrescu, A.: Evolution Computation, 21–37 (2000)

Software Risk Assessment: A Review on Small and Medium Software Projects

Abdullahi Mohamud Sharif and Shuib Basri

Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
saakuut@gmail.com, shuib_basri@petronas.com.my
http://www.utp.edu.my

Abstract. Software risk assessment is a process of identifying, analyzing, and prioritizing risks. In general, there are large, medium, and small software projects that each of them can be influenced by a risk. Therefore, it needs a unique assessment process of the possible risks that may cause failure or loss of the project if they occur. In the literature, there are wide range of risk assessment researches conducted toward software projects. But there is at least view researches focusing on risk assessment of small and medium software projects. This creates a gap for the risk assessment research field which can cause most of small and medium project without having risk assessment. Therefore, the main focus of the paper is to give researchers an insight of the current level of risk assessment for small and medium software development projects. Finally, some future directions will be discussed hoping to insight the gap of the risk assessment field for small and medium software development projects.

Keywords: Small and Medium Software Development Projects, Software Risk Assessment.

1 Introduction

Risks are important factor for the development of software projects in this world and by its effects a lot of projects failed. In [5], risk is defined as "*the possibility of suffering loss that describes the impact on the project which could be in the form of poor quality of software solution, increased costs, failure, or delayed completion*". Moreover, all projects share some degree of risk, and most Information Technology (IT) projects have *considerable* risks [6]. Risk can, however, be *reduced* [6], *stewarded* [7], and managed according to tight planning and assessment.

Moreover, according to [8], risk management is divided into risk assessment and risk control. The risk assessment is divided into three sub levels which are risk identification, risk analysis, and risk prioritization. The second part of risk management, risk control, is also divided into risk management planning, risk resolution, and risk monitoring.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 214–224, 2011.

On the other hand, software development projects are divided into large, medium, and small projects which their definition is based on the number of Lines of Code (LOC), duration of the project, and number of developers of the project. In the context of software development projects, small and medium software development projects (SMSDP) are defined as projects that have 50000-100000 LOC [2], 6-12 months, and ten or fewer programmers [3]. Small and medium projects are growing fast in the world as they are taking part in the economic growth of each country. According to [4], *"Small projects typically carry the same or more risk as do large projects.* [While] *many customers and millions of dollars are lost each year on small projects in product and service organizations"*.

From that perspective of risk management and software development classification, we will focus our paper particularly on risk assessment level for small and medium software development projects. On the other hand, The main objective of this review is to give researchers an insight of the current level of risk assessment for SMSDP. Additionally, the paper provides information about the different types of risk assessment models and methods that found in the literature based on the context of risk assessment for SMSDP.

In this paper, research was organized as follows: section 2 gives overview of the review process, section 3 explains current risk assessments in SMSDPs, section 4 presents comprehensive analysis, and finally section 5 summarizes the review.

2 The Review

We have taken different Internet searches to get information on researches toward SMSDP risk assessment. We divided the search into two stages. In the first stage, we have searched risk assessment for SMSDPs only, and the second stage, we have searched software risk assessment without focusing whether its toward small, medium, or large projects. We found a quite number of researches those their focal point was on this domain, but most of them toward large software projects. However, after adept research, we ended up a total of 12 researches on the domain of software risk assessment for both aforementioned stages. Therefore, we have combined the two stage results as we analyzed both of them in their components of SMSDP's focus.

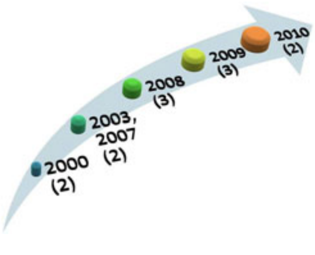


Fig. 1. SMSDP Risk Assesmsnet Timeline

Moreover, the explored researches are in the time span of the last decade. As shown in figure 1, only 3 researches were their center of attention toward software risk assessment in the first half of last decade. Despite the fact that 9 researches are in the direction of software risk assessment consideration in the second half of last decade. That means, as its clear in the picture, the research toward software risk assessment is rising leisurely.

On the other hand, the founded researches was divided based on their proposed outcome into two categories:

- Models category: are those researches provide a process model to assess risk.
- Methods category: are those researches their outcome is method e.g. fuzzy logic method, etc.

Finally, the studied researches with their information of inputs, methods, and outcome will be analyzed and discussed deeply in the following sections.

3 Current Risk Assessments in SMSDP

In this section we divide and analyze each of the aforementioned assessment ways for SMSDPs based on the following models and methods categories.

3.1 Models Category

There is a quite number of models in the literature, which used different procedures or algorithms to assess software risks in general. While some of them prototyped a tool as a proof of concept utilization.

In this section, we summarize the literature of 6 models with their explanation. The explanation includes the model focus, proposes of the model, a brief description of the model, inputs of the proposed model, risk ranking approach of the model, decision analysis taking types of the model, and if the model implements a proof of concept prototype tool. The detailed information for the contribution of each model is summarized in below

Model [1]

- *Focus:* Assessment, treatment, and monitoring automatically risks related in project time management for small and medium software development projects, such as errors in estimating time
- *Proposes:* Risk Assessment Tool (RAT) model
- *Description:* RAT model consists 5 interconnected phases: users, project plan input, risk rules which contains risk ranking matrix, risk conditions, and risk scenarios, risk fetching processes, and risk report. The risk assessment is taken in the early phases of the project
- *Inputs:* Project plan (e.g. Work Breakdown Structure (WBS)) and resources

– *Risk Ranking:* Risks are ranked based on risk rank matrix which contains risk category (1-Unknown, 2-Low, 3-Medium, 4-High, 5-Fatal), probability of occurrence, and risk impact (1-Low, 2-Medium, 3-High). The matrix produces 45 ranks for risk.
– *Decision Taking Types:* Hybrid assessment
– *Prototype:* Implemented a web application prototype.

Model [9]

– *Focus:* Risk assessment and estimation of software projects
– *Proposes:* Software Risk Assessment And Estimation Model (SRAEM)
– *Description:* The model takes inputs to estimate efforts, cost, and risk exposures. Then the risk prioritization and ranking is taken after applying Mission Critical Requirements Stability Risk Metrics (MCRSRM) if there is no changes in the requirements after requirement analysis
– *Inputs:* Measurement, model, and assumption errors using the concept of Function point
– *Risk Ranking:* The estimation and ranking risks is done by using two methods: probability by using risk exposure, and software metrics of risk management based on MCRSRM
– *Decision Taking Types:* Quantitative assessment
– *Prototype:* —

Model [10]

– *Focus:* Risk assessment of software projects
– *Proposes:* Software risk assessment model
– *Description:* The model is based on Grey Theory using Analytic Hierarchy Process (AHP) method and entropy method. In the result of the assessment, the author suggests to study further to determine the major software risk factors
– *Inputs:* Risk of demand analysis, project quality, project schedule, project circumstance, technology and project personnel.
– *Risk Ranking:* In weighting of risk index, the research uses a combination of two methods: subjective method (e.g. AHP), and objective method (e.g. entropy method)
– *Decision Taking Types:* Quantitative assessment
– *Prototype:* —

Model [11]

– *Focus:* Software project risk assessment especially evolutionary prototype software's
– *Proposes:* Risk Assessment Model for Software Prototyping Projects

- *Description:* Addresses the risk assessment issue, introducing metrics and a model that can be integrated with prototyping development processes. The proposed model which uses causal analysis to find the primitive threat factors, provides a way to structure and automate the assessment of risk.
- *Inputs:* Requirements, personal, and complexity metrics
- *Risk Ranking:* —
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* —

Model [12]

- *Focus:* Risk assessment for software projects
- *Proposes:* Software Risk Assessment Model (SRAM)
- *Description:* The model makes use of a comprehensive questionnaire, where a set of questions is carefully chosen with three choices of answers each. The answers are arranged in increasing order of risk.
- *Inputs:* Complexity of software, staff, targeted reliability, product requirements, method of estimation, method of monitoring, development process adopted, usability of software, and tools used for development
- *Risk Ranking:* Assigning different weights to the probabilities level of risk of the project according to the impact of the associated risk elements on quality, schedule and cost respectively
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* —

Model [13]

- *Focus:* Software project risk assessment
- *Proposes:* Software project risk assessment model
- *Description:* The model contains risk probability assessment model and risk impact assessment model which includes assessment of loss and comprehensive assessment of risk impact.
- *Inputs:* Risk factor nodes
- *Risk Ranking:* Using conditional probability distribution table (CPT) with risk semantic reduction matrix
- *Decision Taking Types:* Hybrid assessment
- *Prototype:* —

3.2 Methods Category

Common software project risk assessment methods are AHP, fuzzy math method, Delphi method, etc. In details, we summarized below the literature of 6 method with the explanation. The explanation includes the method focus, proposes of the method, a brief description of the method, inputs of the proposed method, risk ranking approach of the method, decision analysis taking types of the method, and if the method implemented a proof of concept prototype. The detailed information for the contribution of each method is summarized in below.

Method [14]

- *Focus:* Cost and quality of software projects
- *Proposes:* Expectation-Maximization (EM) algorithm
- *Description:* the algorithm enhances the ability in producing hidden nodes caused by variant software projects
- *Inputs:* The probability vector of the top-level nodes
- *Risk Ranking:* —
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* Assessment Tool

Method [15]

- *Focus:* Software risk assessment
- *Proposes:* Source-based software risk assessment method
- *Description:* The method takes into account primary facts based on workshop and secondary facts which a framework is developed.
- *Inputs:* Secondary fact retrieval taken from organization through interviews with stakeholders, and primary fact retrieval which is analyzed from the source of the system
- *Risk Ranking:* —
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* —

Method [16]

- *Focus:* General software development but its only for risk identification
- *Proposes:* A concrete implementation method of risk identification based on the improved Kepner-Tregoe Program
- *Description:* Kepner-Tregoe program uses 4 analysis methods: Problem analysis (PA), Decision analysis (DA), Potential Problem analysis (PPA), and Situation analysis (SA). Each of them differs in objectives and also in application procedure respectively. Therefore, the authors' selected PPA for their risk identification as it's a kind of checklist method.
- *Inputs:* Checking vulnerable areas of the project along the extended vulnerable areas
- *Risk Ranking:* —
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* —

Method [17]

- *Focus:* Risk assessment of software projects
- *Proposes:* Fuzzy expert system
- *Description:* The system includes expertise to evaluate risk of software projects in all respects by using Fuzzy inference

- *Inputs:* Corporate environment, sponsorship/ownership, relation ship management, project management, scope, requirements, funding, scheduling & planning, development process, personnel & staffing, technology, and external dependencies variables
- *Risk Ranking:* Risk matrix based on probability and severity measurements
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* Risk assessment fuzzy expert system

Method [18]

- *Focus:* Software project risk assessment
- *Proposes:* Fuzzy linguistic multiple attribute decision making method
- *Description:* The method estimates risk criteria values using linguistic terms based on triangular fuzzy number, and aggregates risk criteria values by multiple attributes decision making
- *Inputs:* Information from experts
- *Risk Ranking:* Risk assessment criterion is used which contains probability, loss, not controllability, and occurrence time. So the risks which have high in all criterion have high priority
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* Case study application for historic data of completed similar projects

Method [19]

- *Focus:* Software risk assessment
- *Proposes:* Risk assessment method
- *Description:* Develops software risk assessment tool using probabilistic interface model based on water fall model
- *Inputs:* Interview-based risk assessment
- *Risk Ranking:* Increasing order of risk by only providing 3 choices. The first choice will contribute 1 mark, 2 marks for the second choice and 3 marks for the last choice
- *Decision Taking Types:* Quantitative assessment
- *Prototype:* Risk Assessment Visualization Tool (RAVT)

4 Analysis of SMSDP Risk Assessments

4.1 Analysis Based on Assessment Parameters

In the previous section, we have grouped different models and methods according to 7 parameters. these parameter are focus, proposes, description, inputs, risk ranking, decision taking types, and prototype. The description parameter, which summarizes the article and decision taking types parameter, which analyzed in section 4.3, will not be analyzed in this section. In this section we will analyze the aforementioned models and methods based on each parameter.

Focus: All articles are focused on risk assessment for software development projects in general. There are some of the articles specified certain scopes under project management areas or under software development methodology. Also there is an article focused on one part of risk assessment branches. on the other hand, there is an article focused on software risk assessment with additional area.

For those focused on software risk assessment with specific scope under project management are article [1] and [14]. Risk related in project time management such as errors in estimating time is focused by [1], while [14] focuses on risks related on cost and quality of software projects. On the other hand, [11] specifically focuses on risk related on evolutionary prototype software's.

More over, article [16] focuses on one of the three branches of risk assessment, that is, risk identification. The estimation of software projects is also focused additionally in article [9].

Proposes: For articles under model's category, they all of them propose models for their risk assessment procedure. While for method's category, they proposed also different methods based on different algorithms. Some of these articles used fuzzy for their proposed methods like [17] and [18], Expectation-Maximization (EM) algorithm like [14], source code based analysis like [15], and concrete implementation method of risk identification based on the improved Kepner-Tregoe Program such as in [16].

Inputs: All articles used different inputs for their risk identification, analyzation, and prioritization process. For models they used different inputs for their risk assessment model, and for methods, they created different methods based on their followed algorithm to assess risks. For detailed information, please refer section 3.1 for models category and 3.2 for methods category inputs.

Risk ranking: Every model or methods has declared specific ranking procedure for the risk, while some does not. For detailed information, please refer section 3.1 for models category and 3.2 for methods category risk rankings.

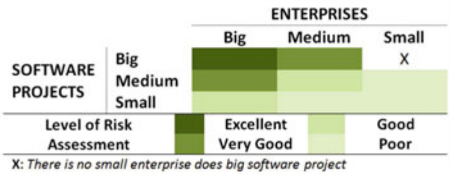
Prototype: For model category, only one article has developed proof of concept prototype for their risk assessment model. Article [1] provides web application prototype using Oracle Application Express (Apex) 3.2 as web tool and Oracle Database 11g as a database tool.

On the other hand, articles [14], [17], and [19] have developed tools or expert systems for their risk assessment methods in the method's category. While [18] takes case study application for historic data of completed similar projects.

4.2 Level of Risk Awareness

There is few researches taken toward small and medium software development project (SMSDP) risk assessment, but most of them is based on a specific aspect of risks, for example, assessing risk in time management of the project [1], or assessing quality risks of the project [14].

In the aforementioned methods and models, almost all risk assessment for software development projects are based on software projects in general without referring whether its small, medium, or large project. As shown in figure 2, level of risk assessment awareness for large and medium software projects in large enterprises have enough assessment by using different commercial tools and framework. While small software projects does not have enough risk awareness. The more the software project size increases the more risk awareness is taken by the enterprises, and the more the software project size decreases the less risk awareness is applied.



X: *There is no small enterprise does big software project*

Fig. 2. Software vs. Enterprise Risk Assessment

4.3 Risk Assessment Decision Taking Types

Taking decision on a risk is based on qualitative assessment, quantitative assessment, or hybrid assessment results. Qualitative assessment means the information are in verbal form rather than in a number or quantity form as in the case of quantitative analysis. Hybrid analysis is combination of both quantitative and qualitative analysis. On the other hand, a survey done by [20] for 10 risk assessment methods, only one method is used qualitative assessment, and another one for hybrid assessment, while the remaining used quantitative assessment.

Based on the aforementioned models and methods in the literature, the decision taking types of them is illustrated in table 1. The main summary that can be made from the table is that the most common type of information that the software risk assessment use is quantitative and in only two cases are used hybrid assessment.

Table 1. DecsionTaking Types

Decision Taking Types	**Model/Method**	**Total**
Quantitative Assessment	[9], [10], [11], [12], [14], [15], [16], [17], [18], [13]	10
Qualitative Assessment	—	0
Hybrid Assessment	[1], [13]	2
Total		12

4.4 Some of the Limitations

The different models and methods mentioned above have some limitations including:

1. The parameters and inputs that each model or method takes are not all of them available in SMSDPs
2. While SMSDPs are rapid development projects and they run from cost, they do not have time to fill all the conditions that methods or models defines

5 Conclusion and Future Directions

We have discussed and analyzed the existing software risk assessment in the literature for the last decades. A total of 12 articles were studied in this paper based on two categories: models and methods. With each category, we examined the articles according into 7 parameters. As we also discussed these parameter in each, based on their different models and methods.

On the other hand, we spotlighted the gap of SMSDP risk assessment in the research field, while we are encouraging other researchers to make their focal point in the direction of SMSDP risk assessment. By the way, solving the abovementioned problems needs different directions. Firstly, this field needs deep research to find the needs and requirement of SMSDPs. Doing brain storming researches are not only enough to fill the gab of the SMSDP needs and requirements, therefore researchers should also focus on the real SMSDP projects to know exactly what those projects requires. Secondly, apart of finding the needs and requirements of SMSDPs, researchers should find also and categories risk factors for SMSDPs locally and globally. This will help to know risk factor of different projects globally. Thirdly, finding factors and requirements of SMSDPs will make easy for other researchers to prepare methods, models, or frameworks that provide suitable approaches for risk assessment of SMSDPs.

Finally, the review taken in this paper is hopefully could give the overall benefits to all researchers in the field of risk assessment for software development projects.

Acknowledgments. The authors would like to thank Universiti Teknologi PETRONAS (UTP) for their cooperation toward submitting and presenting this paper.

References

1. Sharif, A.M., Rozan, M.Z.A.: Design and Implementation of Project Time Management Risk Assessment Tool for SME Projects using Oracle Application Express. World Academy of Science, Engineering, and Technology (WASET) 65, 1221–1226 (2010)
2. Dennis, A., Wixom, B.H., Tegarden, D.P.: System Analysis and Design with UML: An Object-Oriented Approach. John Wiley and Sons, Inc., Chichester (2005)

3. Johnson, D.L.: Risk Management and the Small Software Project. LOGOS International, Inc. (2006)
4. Gray, C.F., Larson, E.W.: Project Management: The Managerial Process. McGraw-Hill Irwin, New York (2008)
5. Boban, M., Poågai, Z., Sertic, H.: Strategies for Successful Software Development Risk Management. Management 8, 77–91 (2003)
6. Brandon, D.: Project Management for Modern Information Systems, pp. 417(6). Idea Group Inc., USA (2006)
7. Barkley, B.T.: Project Risk Management. McGraw-Hill, New York (2004)
8. Boehm, B.W.: Software Risk Management: Principles and Practices. IEEE Software 8(1), 32–41 (1991)
9. Gupta, D., Sadiq, M.: Software Risk Assessment and Estimation Model. In: International Conference on Computer Science and Information Technology, ICCSIT 2008, pp. 963–967 (2008)
10. Qinghua, P.: A Model of Risk Assessment of Software Project Based on Grey Theory. In: 4th International Conference on Computer Science Education, ICCSE 2009, pp. 538–541 (2009)
11. Nogueira, J., Luqi, Bhattacharya, S.: A Risk Assessment Model for Software Prototyping Projects. In: Proceedings of 11th International Workshop on Rapid System Prototyping, RSP 2000, pp. 28–33 (2000)
12. Foo, S.-W., Muruganantham, A.: Software risk assessment model. Management of Innovation and Technology. In: Proceedings of the 2000 IEEE International Conference on ICMIT 2000, vol. 2, pp. 536–544 (2000)
13. Tang, A.-g., Wang, R.-l.: Software Project Risk Assessment Model Based on Fuzzy Theory. In: International Conference on Computer and Communication Technologies in Agriculture Engineering, pp. 328–330 (2010)
14. Yong, H., Juhua, C., Huang, J., Liu, M., Xie, K.: Analyzing Software System Quality Risk Using Bayesian Belief Network. In: IEEE International Conference on Granular Computing, GRC 2007, p. 93 (2007)
15. van Deursen, A., Kuipers, T.: Source-Based Software Risk Assessment. Software Maintenance. In: Proceedings of the International Conference on ICSM 2003, pp. 385–388. IEEE Computer Society, Los Alamitos (2003)
16. Nagashima, T., Nakamura, K., Shirakawa, K., Komiya, S.: A Proposal of Risk Identification Based on the Improved Kepner-Tregoe Program and its Evaluation. International Journal of Systems Applications, Engineering and Development 4(2), 245–257 (2008)
17. Iranmanesh, S.H., Khodadadi, S.B., Taheri, S.: Risk Assessment of Software Projects Using Fuzzy Interface System, pp. 1149–1154. IEEE, Los Alamitos (2009)
18. Li, Y., Li, N.: Software Project Risk Assessment Based on Fuzzy Linguistic Multiple Attribute Decision Making. In: Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, November 10-12, pp. 1163–1166 (2009)
19. Sanusi, N.M., Mustafa, N.: A visualization tool for risk assessment in software development. In: International Symposium on Information Technology, ITSim 2008, vol. 4, pp. 1–4 (2008)
20. Georgieva, K., Farooq, A., Dumke, R.R.: Analysis of the Risk Assessment Methods - A survey, pp. 76–86. Springer, Heidelberg (2009)

Building Knowledge Representation for Multiple Documents Using Semantic Skolem Indexing

Kasturi Dewi Varathan[1], Tengku Mohd. Tengku Sembok[1],
Rabiah Abdul Kadir[2], and Nazlia Omar[1]

[1] Faculty of Information Science & IT, National University of Malaysia, Bangi,
43600 Selangor, Malaysia
kasturi@um.edu.my, {tmts,no}@ftsm.ukm.my
[2] Faculty of Computer Science & IT, University Putra Malaysia, Serdang,
43400 Selangor, Malaysia
rabiah@fsktm.upm.edu.my

Abstract. The rapid growth of digital data and users' information needs have made the demands for automatic indexing to become more important than before. Indexing based on keyword has proven to be unsuccessful to cater for the current needs. Thus, this paper presents a new approach in creating semantic skolem indexing for multiple documents that automatically index all the documents into single knowledge representation. The skolem indexing matrix will then be incorporated in question answering system to retrieve the answer for users query.

Keywords: skolem clauses; skolem indexing; semantic indexing; question answering.

1 Introduction

Document representation is key point in any IR/QA system. The goal of every information retrieval system is to obtain the most accurate result. In achieving this goal, the system has to solely depend on the knowledge representation that represents the knowledge from each of the documents and have integrated it successfully. It has been proven that good knowledge representation will deliver good retrieval results.

Since we are facing with incredible rate of growing corpus and knowledge resources in digital form, the demands for automatic indexing of these knowledge resources has become very crucial. In dealing with multiple documents, many questions arise on how to deal with the inconsistencies issues of the knowledge base and how to integrate these documents into single representation. Thus, this research has focused on how to create a semantic matrix index that represents multiple documents. This indexing will then be used in retrieving the result for users query and also gives proof on which document the answer has been extracted.

2 Literature Review

Creating single knowledge representation from multiple documents research is not something new. There are many researches that have been conducted in dealing with

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 225–231, 2011.

multiple documents [1]. Initial IR methods were based on keywords and Boolean queries. Existing indexing techniques that is widely used by search engine is vector space model in which the indexing is done using keywords.

It is doubtful that these IR methods will be useful in semantic indexing due to lack of precision. The main reason for this drawback is that these systems focused on bare lexical skeleton and leaves out all the linguistically relevant information [2,3]. The consequence of using keyword based indexing is that the irrelevant information that uses certain word but in different context might be retrieved or relevant information which has been represented in a different way might be missed [3,4,5].

User's information needs will not be satisfied with indexing which is based on keywords alone. An alternative way to go beyond bag of words is to organize indexing terms into a more complex structure. With more semantic information about the document captured, it enhances the performance automatically in which higher precision can be achieved by indexing semantic representation rather than keywords [5]. Many researchers currently have and continue to work on semantic representation of documents [6]. There are researches in which text had been indexed by using semantic relation between concepts that reflect the meaning rather than just words [7]. As for [8], has proved that utilizing semantic relation by using wordnet has improved the effectiveness of IR systems. Besides that, thematic relationship between parts of text using a linguistic theory called Rhetorical Structure Theory(RST) have also being indexed and used as a knowledge representation for effective information retrieval [5]. As for [9], they have used the logical structure of a document by utilizing the hierarchy of titles and paragraphs to extract the semantic relations between terms. Meanwhile, [10] has used first order logic representation in performing document indexing for its logical linguistic document retrieval system. On the other hand, [11] has extended [10] and translated the first order logic representation to skolem representation and used skolem representation in indexing single document in their retrieval system. As for our research, we have extended [11] to cater for multiple documents to be indexed as a single knowledge representation.

3 Semantic Skolem Index Creation

Indexing has been an important element that determines the success of text retrieval. In IR, each document is characterized as a set of index terms that exist in the document [12]. And these index terms represents the keywords of the documents. Thus, this representation does not take into account semantics of text documents during the indexing process. In our research, we no longer use keyword to represent as index terms. Instead, we are using skolem clauses to represent the index terms. This skolem indexing managed to tie the associations that exist between the skolem representation and these associations becomes the ultimate information that helps in the retrieval process. Knowing the importance of association, this research will basically show how we have stored the association that lies between the skolem representations. In this section we propose a framework on the semantic skolem index creation as shown in figure 1.

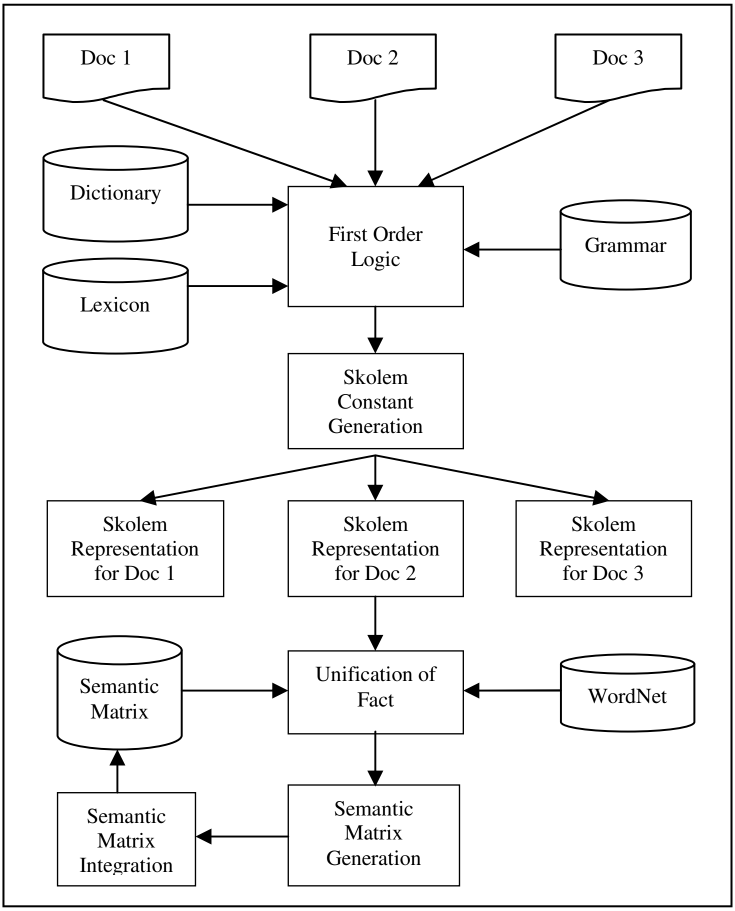

Doc 1
Doc 2
Doc 3
Dictionary
Lexicon
First Order Logic
Grammar
Skolem Constant Generation
Skolem Representation for Doc 1
Skolem Representation for Doc 2
Skolem Representation for Doc 3
Semantic Matrix
Unification of Fact
WordNet
Semantic Matrix Integration
Semantic Matrix Generation

Fig. 1. Semantic Skolem Indexing Framework

We have made used [11] in creating skolem representation for each of the document and have expanded their research to accommodate multiple documents. This framework shows the flow of a document and how it is being represented before the unification of fact process begins. The unification process incorporates WordNet and data which in existence in the semantic matrix database. These data have been incorporated in order to deal with the common problem that exist in indexing multiple documents in which we have to handle similar sentences with more or less the same meaning but have been constructed by using different sets of words. In this kind of circumstances, information needs to be filtered before it gets loaded into the semantic index matrix. For an example, “Marry is a beautiful girl”, and “Marry is a pretty girl”. If “Mary is a beautiful girl” has been represented as skolem representation in semantic index matrix, then the second statement “Mary is a pretty girl” will not be added as new skolem representation in the matrix. But, the frequency of skolem that has been represented earlier will be increased. Thus, the final representation will only have the skolem representation value of “Mary is a beautiful girl” with frequency of existence=2.

The semantic index matrix that has been created is scalable in parallel with the growth of the documents. This scalability feature of the semantic matrix enable the real time environment data in document form to be integrated automatically to it.

Example

Here is an example of small text documents that have been used for illustrative purpose. In this example, each document consists of 1 or 2 sentences. The sentences are represented in natural language.

Document1: Chris has written two famous books.
Document 2: A man named Noah wrote this book.
Document 3: Noah works in school. Noah starts to write this book to give the right spelling.
Document 4: Winnie the Pooh was written in 1925.
Document 5: The process of writing book is very tedious.
Document 6: The author wrote books. The author writes these books to share his views on politics.
Document 7: The author writes books in English.

Table 1. The skolem representation for document 1 to document 7

Document 1:	Document 2:	Document 3:	Document 4:
two(g48). famous(g48). book(g48). ***writes(chris,g48).***	man(g119). book(g116). ***writes(names(g119,noah),g116).***	school(g14). ***works(noah,g14).*** book(g15). ***writes(starts(noah), g15).*** right(g18). ***gives(writes(starts(noah),g15),g18).*** spelling(g19). ***gives(writes(starts(noah),g15),g19).***	***writes(r(winnie & pooh),1925).***
Document 5:	Document 6:	Document 7:	
process(g5). writes(g2). ***of(g5,g2).*** book(g6). tedious(g7). ***isa(g6,g7).***	author(g40). book(g41). ***writes(g40,g41).*** author(g40). book(g41). ***writes(g40,g41).*** his(g46). view(g46). politic(g47). ***on(g46,g47).*** ***shares(writes(g40,g41),g46).***	author(g30). book(g31). ***writes(g30,g31).*** ***in(writes(g30,g31), english).***	

The skolem representation for each of the documents have to be unified and integrated into single knowledge representation in order to be used in semantic retrieval process. We assume that $S=\{s_1,s_2\ldots s_n\}$ is the set of all unified skolems used in indexing the documents form $D=\{d_1,d_2,\ldots..d_n\}$. A document can be seen as a set of skolem representation, that is, $d_i=\{s_1,s_2,\ldots s_k\}$ where s_j denotes the skolem j in document d_i.

The knowledge representation that has been proposed in this paper has a single knowledge representation that represents the whole documents that will be generated instead of using sub knowledge representation from each of the documents. For indexing purpose, we have build a skolem-document matrix in which the rows represent the number of occurrences of all the unified skolem constants in the corpus and the column represents the document numbers.

4 Experimental System Development

The translation process that translated text documents to first order logic is being done by using prolog. We have used pragmatic skolemization technique by using prolog as a tool in translating first order logic to skolem representation. The process of indexing the skolem clauses has been done using php and the index has been stored in mysql database. In retrieving the answer for users query, resolution theorem proving approach has been used [13]. The question will be used as a theorem to be proven in order to derive to the answer which has been stored in the skolem-document index matrix. Skolem clause binding approach as in [11, 14] have been used to bind all the interrelated skolems together that is bound by the answer key.

5 Experimental Result and Discussion

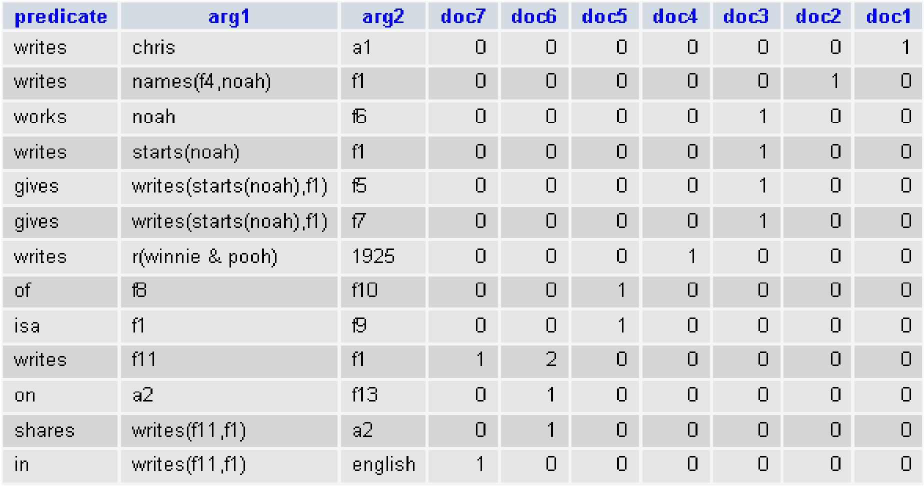
The result in figure 2 shows the skolem-document matrix where rows represent all the possible skolem clauses and the column represents all the documents. The unified skolem that has 2 arguments are listed in the figure as shown above. We took advantage of the association between each of the skolem representation from multiple documents in retrieving accurate answer for our question answering system.

As for [15] has dealt with matching the query content with available documents together with the most appropriate fragments of this document. On the other hand, [14] managed to retrieve the exact answer in logic representation from single documents. But we have gone a step ahead in which providing the user with the answer for his/her queries from multiple documents and retrieving the documents in which the answer contains as proof for the respected queries posed [13].

Here is an example of query posed by the user.

Query: *Who writes book?*

writes(chris,a1)	Doc no 1=1
writes(names(f4,noah),f1)	Doc no 2=1
writes(starts(noah),f1)	Doc no 3=1
writes(f11,f1)	Doc no 6=2, Doc no 7=1



predicate	arg1	arg2	doc7	doc6	doc5	doc4	doc3	doc2	doc1
writes	chris	a1	0	0	0	0	0	0	1
writes	names(f4,noah)	f1	0	0	0	0	0	1	0
works	noah	f6	0	0	0	0	1	0	0
writes	starts(noah)	f1	0	0	0	0	1	0	0
gives	writes(starts(noah),f1)	f5	0	0	0	0	1	0	0
gives	writes(starts(noah),f1)	f7	0	0	0	0	1	0	0
writes	r(winnie & pooh)	1925	0	0	0	1	0	0	0
of	f8	f10	0	0	1	0	0	0	0
isa	f1	f9	0	0	1	0	0	0	0
writes	f11	f1	1	2	0	0	0	0	0
on	a2	f13	0	1	0	0	0	0	0
shares	writes(f11,f1)	a2	0	1	0	0	0	0	0
in	writes(f11,f1)	english	1	0	0	0	0	0	0

Fig. 2. Skolem-Document Indexing Matrix

Set of Skolem Clauses(Answer):
chris
names(man,noah)
noah
author

Answers have been retrieved together with the proof in which document the answer contains and the frequency of occurrence of each of the semantic relation has been successfully retrieved from the semantic index matrix that we have created.

6 Conclusion

The information capturing, semantic generation and semantic integration could involve some preprocessing time during indexing. However these tediousness have been compensated with higher precision in terms of retrieval. The indexing and retrieval technique described in this paper is under development for huge collection of documents, thus we have used small experiments to prove that the method and the retrieval harvested good result.

Acknowledgement

The authors wish to thank the Ministry of Higher Education Malaysia, University of Malaya and the Knowledge Technology group (Center for AI Technology, FTSM, UKM) for assisting and giving the opportunity for the authors to conduct this research.

References

1. Clark, P., Thompson, J.: A Study of Machine Reading from Multiple Texts. In: AAAI Spring Symposium on Learning by Reading and Learning to Read (2009)
2. Hess, M.: Deduction over Mixed-Level Logic Representations for Text Passage Retrieval. In: International Conference on Tools with Artificial Intelligence (TAI 1996), Toulouse, France, pp. 383–390 (1999)
3. Shaban, K.B., Basir, O.A., Kamel, M.S.: Document Mining Based on Semantic Understanding of Text. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) CIARP 2006. LNCS, vol. 4225, pp. 834–843. Springer, Heidelberg (2006)
4. Egozi, O.: Concept-Based Information Retrieval using Explicit Semantic Analysis, pp. 1–80. Master of Science in Computer Science: Israel Institute of Technology (2009)
5. Marir, F., Haouam, K.: Rhetorical Structure Theory for content-based indexing and retrieval of Web documents. In: 2nd International Conference on Information Technology: Research and Education-ITRE 2004, London, pp. 160–164 (2004)
6. Hoenkamp, E., van Dijk, S.: A Fingerprinting Technique for Evaluating Semantics Based Indexing. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 397–406. Springer, Heidelberg (2006)
7. Zambach, S.: A Formal Framework on the Semantics of Regulatory Relations and Their Presence as Verbs in Biomedical Texts (2009)
8. Ceglarek, D., Rutkowski, W.: 19 Automated Acquisition of Semantic Relations for Information Retrieval Systems. Technologies for Business Information Systems, 217–228 (2007)
9. Bounhas, I., Slimani, Y.: A hierarchical Approach for Semi-Structured Document Indexing and Terminology Extraction. In: International Conference on Information Retrieval and Knowledge Management (2010)
10. Tengku Sembok, T.M.: A simple logical-linguistic document retrieval system. Information Processing and Management 26(1), 111–134 (1990)
11. Abdul Kadir, R., Tengku Sembok, T.M., Halimah, B.Z.: Towards Skolemize Clauses Binding for Reasoning in Inference Engine. In: Fifth International Conference on Computational Science and Applications (2007)
12. Cai, D., van Rijsbergen, C.J.: Semantic Relations and Information Discovery. SCI, vol. 5, pp. 79–102 (2005)
13. Varathan, K.D., Tengku Sembok, T.M., Abdul Kadir, R., Omar, N.: Retrieving Answer from Multiple Documents Using Skolem Indexing. In: International Conference on Semantic Technology and Information Retrieval (2011) (in press)
14. Abdul Kadir, R., Tengku Sembok, T.M., Halimah, B.Z.: Improvement of document understanding ability through the notion of answer literal expansion in logical-linguistic approach. WSEAS Transactions on Information Science and Applications 6(6), 966–975 (2009)
15. Prince, V., Labadí, A.: Text segmentation based on document understanding for information retrievalp. In: Applications of Natural Language to Data Bases, pp. 295–30 (2007)

Synchronous Replication: Novel Strategy of Software Persistence Layer Supporting Heterogeneous System

A.H. Beg[1], Noraziah Ahmad[1], Ahmed N Abd Alla[2],
E.I. Sultan[1], and Noriyani Mohd Zin[1]

[1] Faculty of Computer Systems & Software Engineering
[2] Faculty of Electric & Electronics Engineering
University Malaysia Pahang
Gambang-26300, Pahang, Malaysia
ahbeg_diu@yahoo.com, {noraziah,ahmed}@ump.edu.my

Abstract. Synchronous replication is the ideal solution for organizations: search for the fastest possible data recovery, minimal data loss and protection against the problems of database integrity. This ensures that the remote copy of data that is identical to the primary copy is created at the same time the primary copy is updated. However, most of the synchronous replication does not consider the heterogeneous system. In this paper, a software persistence layer for heterogeneous synchronous replication has been designed and developed based on multi-threading known as PLSR. The main objective of this strategy is to make the persistence layer adaptive and make the synchronous replication process reliable and faster than other existing replication processes concerning cost minimization. In the proposed PLSR replication technique, the replication servers are OS independent and the entire replication process is not inter dependent nevertheless on the main server. Adding a new replication server is easier than other processes. The technique also introduces the modification of replication servers without making impairment to the entire process. The comparative Results with SQL server data replication show the PLSR is more acceptable in terms of transactional insertions and sync time. The result shows that PLSR outstanding performs 88.5 % faster than SQL server for transactional insert.

Keywords: Data replication, Synchronous, Heterogeneous replication, Multi-threading technique, Software persistence layer.

1 Introduction

Data replication is the process that maintains multiple copies of data, called replicas, on separate computer. It can improve the availability by allowing access to the data even when some of the replicas are unavailable. Replication also can improve the performance by the following: i) reduce latency, since users can access nearby replicas. Therefore, avoiding remote network access, ii) increasing throughput, since the multiple computer is potential to serve the data simultaneously [1]. Replication can be used to enhance availability and performance in distributed systems. In the data-centric distributed system, replica consistency and data integrity (constraint consistency) are used as correctness criteria [2].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 232–243, 2011.

The main objective of replicas is to increase the system reliability and application performance. In the grid community, distributed and clustering system lot of work has focused on providing efficient and safe replication management services through designing of algorithms and systems. Businesses or specially Enterprise business or industrial business use replication for many reasons. Replication technology creates data replication on the right node from where the data transmission becomes faster. Like a network is in some remote location separated from the main server, and the data transmission rate is too high. Thus a replication server can be created on that remote location which in terms helps the remote system reduce data transmission impediments and improve visit delay, bandwidth consumption and system reliability [3].

In the grid environment, Hitoshi Sato et al. [4] proposed an approach for the clustering base replication algorithm. The goal is to create a technique to automatically determine optimal file replication strategies. Their approach outperformed groups file stored in a grid file system according to the relationship of simultaneous file access and determines locations and movement of replicas of file clusters from the observed performance data of file access and implementation specification was in Linux 2.6.1.8. The authors do not consider the heterogeneous system and also the replication in the grid environment needs a lot of inter connection speed (gigabyte). Ali Elghirani et al. [5] proposed an approach in an intelligent replication framework for data grid. The main goal of their approach is to create a replica management service that interrogates replica placement optimization mechanisms and dynamic replication techniques, coupled with computation and job scheduling algorithms for better performance in data grids. They use dynamic ordinary replication strategies and replica placement schemes. Their result shows that their approach improves the job execution time by 10-23%. Their work, however, replica management is only coupled with computational job scheduling, which actually better performs in Symmetric Multi-Processors Server (SMP).

Yuan - sheng Lou et al. [6] studied a reflected persistence data layer framework based on O/R mapping was designed and implemented. Persistence data layer is the most important part in the information system design. It is the foundation of the performance of the system and its migration ability. In this paper, they presented five modules: data loadable module, data write a module, database services module, primary key cache module and paging cache module for persistence layer. However, reflection is not native to the OS. A lot of execution handling mechanisms should be included into the system. Besides replication using the reflection mechanism is a very slow process and takes a lot of memory and might cause the buffer overflow.

In the peer-to-peer network using dynamic replication proposed a load sharing technique [7] providing and improving access performance there has been proposed two load sharing techniques, which use data replication. At the first technique there has been used a periodic push-based replication (PPR) to reduce the hop count (the number of legs traversed by a packet) and at the second technique it uses on demand replication (ODR) that performs and improves access frequency. However, they proposed two algorithms: improve access performance on a P2P network.

In the current enterprise software system, there used a persistence layer which persists in different current objects, which in terms help the application to avoid fault tolerance. Data replications used in the different field, such as the bank, insurance; group of industries to help protects their secure data to avoid any unwanted crashes. Data replication in terms of duplication of data creates a backup copy of the data on the different server. So basically, on an enterprise system replication helps to avoid fault of the data server system. When data insertion, deletion or any modifications happens to the main database server it in term reflexes to the replicated server. Currently, data replication system and software development practice have the following dependencies and/or impediments:

i. Usually replication process depends on the main server.
ii. It is hard to make a decision when a replicated database server crash.
iii. Introducing the up gradation of the replication process usually mutes or pause the system for a routine of time.
iv. Fail or cashes of the main server usually make the entire system stop working (For a database driven system)
v. Managing replication servers are cost effective.

The main contribution of this work is to design and development of software persistence layer for synchronous replication that support the heterogeneous system. The proposed layer used as a multi-threaded application and which also provides an interface between the database and the main system. The persistence layer has a single thread which is responsible for making communication with the main server and has another thread running to manage the replicated databases. So it helps the entire system to reduce dependency of the replicated server on the main server. Replication server also used as a main server in the case of the crashing or fail of main server. So adding more replicated servers are alike plug and play features. Finally, in the result section, results and discussion of the proposed replication system (PLSR) compares with the replication process of SQL Server (transactional and merge insert time).

2 Background

Replicated database and a distributed database sound the same. However, in a distributed database, data is available at many locations, but a particular table resides at only one location [8]. For example, the employee's table resides at only the pah.employee database in a distributed database system that also includes the kl.employee and kn.world databases. Nevertheless, replication means that the 100% same data at another location [9]. Replication balances the data transaction, and it provides fast, local access to shared data over multiple sites [10]. So replication works as a load balancing.

Data replication can be drives by programs which transport data to some other location and then loaded at the receiving location. Data may be filtered and transformed during replication. Replication must not interfere with existing applications and should have the minimal impact on production systems. The replication processes to need to be managed and monitored [11]. Therefore, data replication improves data access time, transaction time and provides fault tolerance by maintaining and managing multiple copies of data (e.g. files, objects, databases or parts of databases) at different locations [12]. A replication environment can use either asynchronous or synchronous to copy data. With asynchronous replication, changes are made one after a certain time with a lot of data from the master site to the different other site. With synchronous replication, changes made immediately once some data transaction occurs to the mater site. Using synchronous replication, an update of transaction results eventually replication of the update at all other sites [13].

The main benefit of synchronous replication is that data can be retrieved quickly. Operations on the remote mirror site may begin immediately when the primary site should be stopped working on the primary site to be disturbed. Few actions in the process at the time of failure may be lost. Because neither primary nor the remote sites have a record of these transactions, the database rolls back to the last state confirmed [14].

2.1 Heterogeneous System

A distributed heterogeneous computing system is a collection of autonomous dissimilar computing machines that are linked by a network and are coordinated with software functioning as a single powerful computing facility. Heterogeneous system can provide low cost and high performance computing whenever computational applications can be broken into tasks that can be distributed the various machines for parallel execution. A Distributed Heterogeneous Computing system has potential advantages than the homogenous system because some tasks run faster on one type of machine while other types of tasks may run faster on the different machine [15].

The heterogeneous computing system is the very talented platform because the single parallel architecture based system might not be sufficient for running application to exploit the parallelism. Sometimes, heterogeneous distributed computing systems can achieve higher performance than single super computer systems; furthermore, it puts the lower cost than the super computer. Conversely, the HDC system is more exceptions oriented so it might put the negative impact on the running application [16]. The homogenous computing is the system is easier to control because the processing time is independent and identically with an arbitrary identical distribution [17]. The heterogeneous Computing systems can achieve both capability and capacity based jobs where capability based (aimed at minimization of the completion time of one big job) and capacity based (aimed at maximization the number of completions of small jobs within a given time) [18].

2.2 Persistence Layer

Persistence layer (Layer Architecture) provides an abstract interface for data access layer that is part of a storage mechanism (s). This type of interface is abstract and independent of storage technology. Typical features include:

i. Store and/or Retrieve of the whole database objects
ii. Abstraction of the database cursor with all instances of a given type
iii. All available transaction support, including open, commit, abort and rollback
iv. Data session management
v. Data Querying support.

Normally, the persistence layer is the construction of at least two inner layers of an application: the first includes an abstract interface and the second is a set of binding to each target database. The implementation may have more than two internal divisions between the logic layer and the storage mechanism layer [19].

2.3 Transactions

A transaction is a group containing a set of tasks which inherent part of any application that collects or manipulates data. SQL server has to make sure the data integrity. That means two users should not modify the same piece of data at the same time. In the SQL server, a transaction is a single statement or multiple data modification language (DML) statements executed together. In a transaction, all statements are treated as a group of works. If one of the statements fails then the transaction treated as a fail transaction and the whole transaction roll back. As a result none of the changes are saved. On the other hand, if all the statements are succeeding the transaction treated as a succeed transaction and committed [20, 21].

3 Complete Flow Chart

In this technique, the system promotes a GUI which facilitates users to insert data to the system. From the GUI, data send to the persistence layer. To configure the servers i.e. to populate the server's information persistence layer check the configuration and connection string file either, which is existed and readable. Exception handler generates messages to send to the user. Alike with configuration file Exception handler covers connection string weather the connection string is readable or not. An unreadable connection string shows a message to the user, other than that the system read connection string where the database connection URL stored as a XML file. The persistence layer creates multithread based on the configuration file, i. e. the definition of main server and the replication servers along with the numbers (number of replication servers). For the main server, persistence layer creates a high priority thread. The high priority thread is responsible for the transactions to the main server. Along with this, for the X number of replication server, persistence layer creates X number of low priority threads, which provide service for the transaction to the replication servers. A notification is sent to the end user/ administrator when the entire process is finished. Fig.1 shows the flow chart of the Persistence layer replication process.

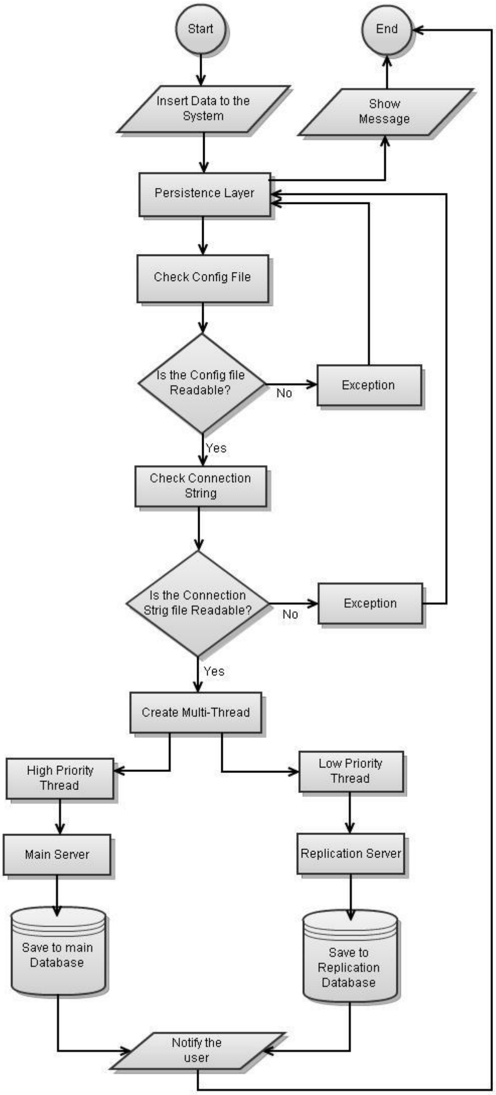


Fig. 1. Flow chart of Persistence layer replication process

4 Proposed PLSR Algorithm

To execute the proposed architecture has been developed different algorithms for different function, which are described below:

4.1 Persistence Layer Algorithm

The persistence layer algorithm parse configuration file and establish the connection among all the replication database servers and the main server. It queues all database transactions for the replication server and the main server.

Algorithm persistence_init(XMLFile:FILE)

```
BEGIN
  Try LOAD XMLfile
  EXCEPTION: Send “FILE NOT FOUND”
  LET $N_S$ as string = NULL
  LET $N_L$ as string = NULL
  WHILE not EOF (XMLFile)
          BEGIN
             READ REP→Serv→MS→Name
             ASSIGN Name to the $N_S$
             READ REP→Serv→MS→Loc
             ASSIGN Loc to the $N_L$
          END
  LET $R_N$ = {} as empty set where $R_N$ represents the name of the replication
  server
  LET $R_L$ = {} as empty set where $R_L$ represents the location of the replication
  server
  WHILE not EOF (XMLfile)
   BEGIN
           PARSE XMLfile
          IF LINE not starts with is “MS”
          ASSIGN value in Rep→Serv→ServerName into $R_N$
          ASSIGN value in Rep→Serv→Location into the $R_L$
  END
 END
```

4.2 Connection String Algorithm

As the name suggests, the connection string basically stores the way to connection to the server. In this system, the connection strings are located into an XML file. Thus, the connection string algorithm parses the connection string from the XML file and sends the string value to the Persistence layer algorithm to establish the connectivity.

Algorithm persistence_read_connection_string(XMLFile:FILE)

```
BEGIN
  Try LOAD XMLfile
  EXCEPTION: Send “FILE NOT FOUND”
  LET $C_s$ as string = NULL
  LET $C_L$ as string = NULL
  WHILE not EOF (XMLFile)
          BEGIN
             READ REP→Serv→ConnectionString
```

```
            ASSIGN ConnectionString to the $C_S$
            READ REP→Serv→OSInfo
            ASSIGN OSInfo to the $C_L$
        END
    LET $S_N$ = {} as empty set where $S_N$ represents the ConnectionString of the
    replication server
    LET $S_L$ = {} as empty set where $S_L$ represents the OSInfo of the replication
    server
    WHILE not EOF (XMLfile)
    BEGIN
        PARSE XMLfile
        IF LINE starts with is "RepServ"
        ASSIGN value in Rep→RepServ→ConnectionString into $S_N$
        ASSIGN value in Rep→RepServ→OSInfo into the $S_L$
    END
END
```

4.3 Lookup Service Algorithm

In this system, lookup service is the special layer from where lots of synchronization occurred. The system keeps track of lowest traffic information into the database and done a routine check that at when it can make synchronization. Synchronization is necessary if the config file got changed or if the new replication server added.

```
Algorithm Data_Synchronization(XMLFile:FILE)

BEGIN
    LET T= Empty where T represents the Main server DATETIME
    LET $T_F$ = Empty where $T_F$ represents the LOW TRAFFIC info from the server
        IF T=$T_F$ Then
        CALL Utility Function start_data_synchronization()
        EXIT
    END IF
    LOAD XMLFile
    LET $T_C$ = NULL where $T_C$ represents the config file created date
    LET $T_M$ = NULL where $T_M$ represents the config file modification date
    LET $T_{LM}$ = Empty where $T_{LM}$ represents the config file's last modification
    date
        IF ($T_C$ =$T_M$ or $T_C$ $T_M$ =$T_{LM}$ ) Then
        CALL Function start_data_synchronization()
    ELSE
    LOAD XML file
    PARSE XML file
    CALL Function persistence_init(XMLFile)
    END
```

4.4 Utility Algorithm (Add Previous Record)

The utility functions show how synchronization happened between the main server and a newly added replication server. The system creates a log.txt file which stores all the SQL commands for all tables and data. After that it executes those SQL commands to the newly added replication server.

```
Algorithm  Add_previous_record(Rep_ServerName)

BEGIN
  CREATE file log.txt
  LET d = {} is a empty set represents all the db tables of the main server
  BEGIN
          READ table names from the d
          WRITE table names into log.txt
          CREATE SQL command to CREATE tables into Rep_Server_Name
          WRITE into log.txt
  END
  WHILE table names is not End
          BEGIN
                    READ data from main server table
                    WRITE data into rep server table
          END
  END while
END
```

4.5 Utility Algorithm (Synchronize Data)

This is another utility function to synchronous data between the main server and replication servers. When any data couldn't replicate to the replication server then it stores to an XML file as a SQL command. After that at the lowest traffic time it parses all the SQL command and executes to the replication server.

```
Algorithm synchronization_data(commandxmlfile:File )

BEGIN
  PARSE command XMLfile
   LET $P_C$= {} empty set represents all the SQL command
  WHILE command XMLfile is not EOF
          BEGIN
           PARSE command from commandXMLfile
                   ASSIGN command into $P_C$
          END
  END while
  WHILE $P_C$ is not end
          BEGIN
                   EXECUTE SQL command
          END
  END while
END
```

4.6 Notation of PLSR Algorithm

Various notions have been used in the PLSR algorithm. The definition of the notation has been described in the Table 1.

Table 1. PLSR Notation and Definition

Variable	Definition
N_S	Main Server Name
N_L	Main server Location
MS	Main server in the XML tag
Rep	Replication server in the XML tag
Loc	Location of the replication server in the XML tag
R_N	The list of replication server name
R_L	The list of replication server location which has been read from XML file
Name	Name represents the name of the server
Location	Location represents the network location of the server
Serv	Serv represent the server
C_S	Main server's connection string
C_L	Main Server operating information
S_N	Replication server's connection string
S_L	Replication server's operating information
T	The current data time of the main server
T_F	The lowest traffic information from the database
T_C	From the config it can find the information that when the file created then it represented by Tc
T_M	From the config file, when the file has been modified then it represented by T_M
T_{LM}	The last modifications date represented by T_{LM}
Pc	Pc represent as the set of SQL comamnd

5 Result and Discussion

This paper compares the execution time with the SQL Server merge and transaction insertion with the PLSR replication. Table 2 shows the comparative result.

Table 2. Comparison between SQL Server and PLSR

No. of Rows	SQL Server Transaction Insert	SQL Server Transaction Synchronization	PLSR Transaction Insert
100	0	0	0.659
500	**1**	**1**	**0.432**
1000	2	1	0.768
5000	7	2	1.987
10000	26	5	2.967

Table 2 demonstrates that, for 500, 1000, 5000 and 10000 data insertions (transactional insertion) SQL Server replication takes 1, 2, 7 and 26 seconds. Conversely, the PLSR replication takes 0.659, 0.432, 0.768 and 2.967 respectively.

The total replication time (RT) has been calculated using equation (1)

$$RT = \sum (TT+ST) \tag{1}$$

Here, RT represents Replication Time; TT represents Transactional Time and ST represents Synchronous Time.

For 500 rows of data insertion in SQL server,
RT = (1+1) Seconds
= 2 Seconds

For 500 rows of data insertion in PLSR,

RT = (0.432+0) Seconds [ST=0, Because PLSR completed TT and ST
= 0.432 Seconds at the same time]

From the replication time, it shows that PLSR replication time is lower than SQL Server replication. As the number of data goes higher, SQL replication time getting much higher in compared to the PLSR replication.

The motivation to compare the result with SQL Server replication is, merge, and transaction insertions can alter using several trigger, which is similar with our strategy, as the algorithm can perform rollback command from the persistence layer which can alter the result. From this above result and the execution point of view, PLSR is more acceptable.

6 Conclusion

This paper has been designed and developed a persistence layer for synchronous replication (PLSR) that supports heterogeneous system. It can help the enterprise application more secure and reliable data transmission. One of the main goals is to make the database replication more and easier to handle thus make it vastly configurable and also the whole architecture is the service oriented means; it used latest technology trends and the replication will be from the persistence layer. Persistence layer is a part of the software engine, and it used the latest customizable fourth generation language like c# or Java. Therefore, a new era can begin related to networking and as well as database programming. The comparative result between the SQL server replication and PLSR replication shows that PLSR replication is more acceptable in terms of transactional insertions and sync time. Further work of this paper will include the implementation of the algorithm which supports synchronous replication and as well as creating a distributed application.

References

[1] Abdul-Wahid, S., Andonie, R., Lemley, J., Schwing, J., Widger, J.: Adaptive Distributed Database Replication Through Colonies of Pogo Ants. In: IPDPS, IEEE International Parallel and Distributed Processing Symposium, pp. 1–8 (2007)

[2] Osrael, J., Froihofer, L., Chlaupek, N., Goeschka, K.M.: Availability and Performance of the Adaptive Voting Replication Protocol. In: ARES, The Second International Conference on Availability, Reliability and Security, pp. 53–60. IEEE Press, Vienna (2007)

[3] Gao, T., Liu, F.: A Dynamic Data Replication Technology in Educational Resource Grid. In: ISITAE, First IEEE International Symposium on Information Technologies and Applications in Education, pp. 287–291 (2007)

[4] Sato, H., Matsuoka, S., Endo, T.: File Clustering Based Replication Algorithm in a Grid Environment. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 204–211 (2009)

[5] Elghirani, A., Zomaya, A.Y., Subrata, R.: An Intelligent Replication Framework for Data Grids. In: AICCSA, IEEE/ACS International Conference on Computer Systems and Applications, pp. 351–358 (2007)

[6] Lou, Y.-s., Wang, Z.-j., Huang, L.d., Yue, L.-l.: The Study of A Reflected Persistence Data Layer Framework. In: WCSE, WRI World Congress on Software Engineering, pp. 291–294 (2009)

[7] Rajasekhar, S., Rong, B., Lai, K.Y., Khalil, I., Tari, Z.: Load Sharing in Peer-to-Peer Networks using Dynamic Replication. In: AINA, 20th International Conference on Advanced Information Networking and Applications, pp. 1011–1016 (2006)

[8] Wang, Y., Li, S.: Research and performance evaluation of data replication technology in distributed storage systems. Computers & Mathematics with Applications 51, 1625–1632 (2006)

[9] Bost, Charron, B., Pedone, F., Schiper, A.: Replication Theory and Practice, ch. 2. Springer, Heidelberg (2009); ISBN-10 3-642-11293-5

[10] Lin, Y.: Practical and consistent database replication, ch.1-2. McGill University Montreal, Quebec (2007)

[11] Gu, L., Budd, L., Caycl, A., Hendricks, C., Purnell, M., Rigdon, C.: Practical Guide to DB2 UDB Data Replication, vol. 8, ch. 1. IBM, Durham (2002)

[12] Ibej, U.C., Slivnik, B., Robic, B.: The complexity of static data replication in data grids. Parallel Computing 31, 900–912 (2005)

[13] Urbano, R.: Oracle Database Advanced Replication. Oracle Corporation, Part No. B10732-01, ch.1 (2003)

[14] Hitachi data system, http://www.hds.co.uk/assets/pdf/sb-synchronous-data-replication.pdf

[15] Boyera, W.F., Hurab, G.S.: Non-evolutionary algorithm for scheduling dependent tasks in distributed heterogeneous computing environments. J. Parallel Distrib. Comput. 65, 1035–1046 (2005)

[16] Tanga, X., Li, K., Li, R., Veeravalli, B.: Reliability-aware scheduling strategy for heterogeneous distributed computing systems. J. Parallel Distrib. Comput. 70, 941–952 (2010)

[17] Tong, X., Shu, W.: An Efficient Dynamic Load Balancing Scheme for Heterogenous Processing System. In: IEEE Conference on Computational Intelligence and Natural Computing, pp. 319–322 (2009)

[18] Guest editorial. Heterogeneous computing. Parallel Computing 31, 649–652 (2005)

[19] Open Ehr, http://www.openehr.org/208-OE.html?branch=1&language=1

[20] Dewald, B., Kline, K.: InformIT.: SQL Server: Transaction and Locking Architecture, http://www.informit.com/articles/article.aspx?p=26657

[21] Poddar, S.: SQL Server Transactions and Error Handling, http://www.codeproject.com/KB/database/sqlservertransactions.aspx

Preserving Data Replication Consistency through ROWA-MSTS

Noraziah Ahmad[1], Roslina Mohd. Sidek[1], Mohammad F.J. Klaib[2], Ainul Azila Che Fauzi[1], Mohd Helmy Abd Wahab[3], and Wan Maseri Wan Mohd[1]

[1] Faculty of Computer Systems & Software Engineering,
University Malaysia Pahang, Pahang, Malaysia
noraziah@ump.edu.my
[2] Faculty of Science and Information Technology, Jadara University, Irbid- Jordan
[3] Faculty of Electric & Electronic, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Abstract. In modern distributed systems, replication receives particular awareness to provide high data availability, reliability and enhance the performance of the system. Replication becomes as significant mechanism since it enables organizations to provide users with admission to current data where and when they need it. Integrated VSFTPD with Read-One-Write-All Monitoring Synchronization Transaction System (ROWA-MSTS) has been developed to monitor data replication and transaction performs in distributed system environment. This paper presents the ROWA-MSTS framework and process flow in order to preserve the data replication consistency. The implementation shows that ROWA-MSTS able to monitor the replicated data distribution while maintaining the data consistency over multiple sites.

1 Introduction

The beginning of amazing advance in the growth of computer and hardware in computer world enables user to access information anytime and anywhere regardless of the geography factor. In modern distributed systems, replication receives particular awareness to provide high data availability, reliability and enhance the performance of the system [1, 2, 3, 4, 5]. Nowadays, the interest in distributed system environment has increasingly demanded due organization needs and the availability of the latest technology. Ensuring efficient access to such a huge network and widely distributed data is a challenge to those who plan, maintain and handle replication and transaction performance in network [1, 2, 4]. The need of replicated data in distributed systems environment is to ensure that any data is backed up whenever emergency occurs. The replicated data will be imitative in different server(s) in the distributed environment. These advantages of replication are vital since it enables organizations to supply users with admission to current data anytime or anywhere even if the users are physically remote [6]. Moreover, it also can reduce access delay, bandwidth consumption [6], fault tolerance [1, 5, 7, 8, 9, 10] and load balancing [9].

Preserving consistency and availability of a certain data at a huge network becomes the issues that still unsolved. Data organization through replication introduces low data consistency and data coherency as more than one replicated copies need to be updated.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 244–253, 2011.

Expensive synchronization mechanisms are needed to maintain the consistency and integrity of data among replicas when changes are made by the transactions [3]. There are many examples of replication schemes in distributed systems [1, 3, 5, 6, 7, 8]. Synchronous replication model deploys quorum to execute the operations with high degree of consistency and ensure serializability. It can be categorized into several schemes, i.e., all-data-to-all-sites (full replication) and some-data-items-to-all-sites and some-data-items-to-some-sites. One of the simplest techniques for managing replicated data through all-data-to-all-sites scheme is Read-One Write-All (ROWA) technique. Read operations on an object are allowed to read any copy, and write operations are required to write all copies of the object [3, 6]. An available copies technique proposed by Bernstein et al. [11] is an enhance version of ROWA technique, in terms of an availability of the write operations. Every read is translated into read of any replica of the data object. Meanwhile, every write is translated into write of all available copies of that data object. Branch Replication Scheme (BRS) goals are to increase the scalability, performance, and fault tolerance [1]. In this model, each replica is collected of a different set of subreplicas structured using a hierarchical topology. BRS deploys some-data-items-to-all-sites replication scheme. Meanwhile, Neighbour Replication Grid Daemon using some-data-items-to-some-sites replication scheme. Data item has been replicated from primary to adjacent neighbours replica [12].

In our previous work, we present the development of Read-One-Write-All Monitoring Synchronization Transaction System (ROWA-MSTS) [13]. It has been developed to monitor data replication and transaction performs in distributed system environment. However, the paper not discusses the framework and process flow of ROWA-MSTS in order to preserve the data consistency.

In this paper, we review replication concept and recall ROWA-MSTS in Section 2. In addition, we also present existing application related to ROWA-MSTS. Section 3 proposed ROWA-MSTS framework and process flow. In Section 4, we implement proposed framework by deploying real time application in distributed systems environment. The conclusion of this paper is presented in the last section.

2 Related Work

2.1 Concept of Replication

Replication is an act of reproducing. It also addresses the management of the complete copying process [17]. In addition, this process involves the sharing information to ensure consistency between redundant resources, as such the software or hardware components. Data replication may take place if the same data is stored in various storage devices. Meanwhile, computation replication occurs when the same computing task is executed many times [5]. For example, a replicated service might be used to control a telephone switch, with the objective of ensure that even if the main controller fails, the backup can take over its functions.

2.2 ROWA-MSTS

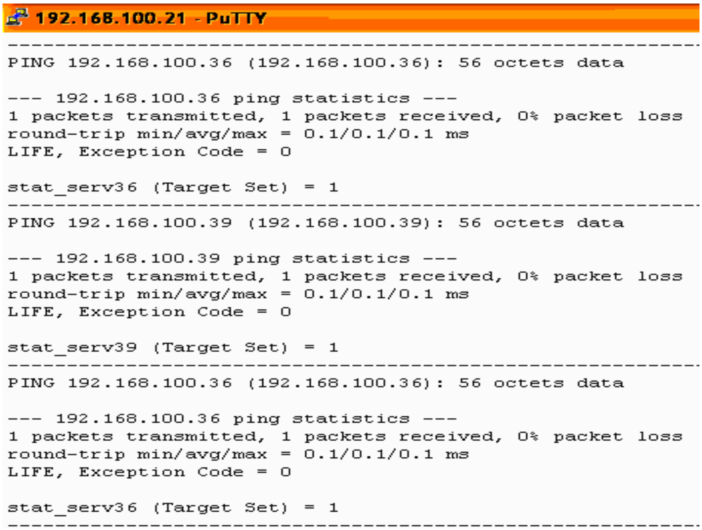
Read-One-Write- All Monitoring Synchronization Transaction System (ROWA-MSTS) [16] has been developed by using ROWA Read-One-Write-All (ROWA) for implementation of data replication management. A read action is allowed to read any copy of data, while a write action is needed to write all copies of data. All of the replicas have the same data when an update transaction commits. All operational sites can communicate with each other. Hence, each operational site can be independently determined which sites are down, simply by attempting to communicate with them. If a site does not respond to a message within the timeout period, then it is assume to be down or unknown status.

ROWA-MSTS integrates with FTP server as an agent communication between replicated servers. From a networking perspective, two main types of FTP includes active and passive. The FTP server initiates a data transfer connection back to the client in active FTP. Meanwhile, the connection is initiated from the FTP client for the passive FTP server. In addition, from a user management perspective also involves two types of FTP. This include regular FTP in which the files are transferred using the username and password of a regular user FTP server, and anonymous FTP in which general access is provided to the FTP server using a well known universal login method. In regular FTP, the VSFTPD package allows regular Linux users to copy files to and from their home directories with an FTP client using their Linux usernames and passwords as their login credentials. VSFTPD also has the option of allowing this type of access to only a group of Linux users, enabling to restrict the addition of new files to system to authorized personnel. ROWA-MSTS integrates with VSFTPD Server. This is because it is fast and stable server. Despite being small for purposes of speed and security, many more complicated FTP setups are achievable with VSFTPD. It is necessary to have a FTP Server install in the entire server since the FTP will make sure the file transfer can be done. After installation of the VSFTPD Server, the server also needs to be configured. ROWA-MSTS was done in shell programming and Bourne Again Shell for the command line editing and jobs control facilities. The job control includes great flexibility in handling the background process. Ubuntu 9.04 Jaunty distribution has been used as the platform for the replicated servers.

2.3 Existing Application Related to ROWA-MSTS

According to Budiarto et. al [5], mobile infrastructure has enabled the introduction of new applications. From business and technology perspectives, data management technology that can support easy access to and from mobile devices is among the main concerns in mobile computing. In implementing the replication strategies, Budiarto et. al assumes that each mobile user holds database which is considered as the master database. A portion of master database on each mobile user is to be shared with other users and therefore become the subject of replication. The replication strategies work in the read-one-write-all (ROWA) context to ensure one-copy serializability.

A.Noraziah et. al [12] developed NRG daemon based on Neighbor Replication on Grid (NRG) Transaction Model. It is one of the tools being used currently where the smart program behaves as agents in distributed systems environment. NRG daemon resolves the timeliness in synchronization by alleviates the lock with small quorum size before capturing updates and commit transaction synchronously to the sites that requires the same update data value. NRG daemon applies the serializability concept during the replication and transaction management. Fig. 1 depicts NRG daemon monitor the activities of the system during the propagation phase.



```
----------------------------------------------------------
PING 192.168.100.36 (192.168.100.36): 56 octets data

--- 192.168.100.36 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.1/0.1/0.1 ms
LIFE, Exception Code = 0

stat_serv36 (Target Set) = 1
----------------------------------------------------------
PING 192.168.100.39 (192.168.100.39): 56 octets data

--- 192.168.100.39 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.1/0.1/0.1 ms
LIFE, Exception Code = 0

stat_serv39 (Target Set) = 1
----------------------------------------------------------
PING 192.168.100.36 (192.168.100.36): 56 octets data

--- 192.168.100.36 ping statistics ---
1 packets transmitted, 1 packets received, 0% packet loss
round-trip min/avg/max = 0.1/0.1/0.1 ms
LIFE, Exception Code = 0

stat_serv36 (Target Set) = 1
----------------------------------------------------------
```

Fig. 1. NRG daemon monitor propagation phases [12]

Yair Amir [18] proposed Postgres as transparent replication structural design. The replication server consists of several autonomous modules that together provide the database integration and consistency services. This includes the Replication Engine, Semantics Optimizer and Postgres specific interceptor. Replication Engine includes all replication logic such as the synchronizer algorithm. It can be applied to any database or application. The engine maintains a consistent state and can recover from a wide range of network and server failures. A Semantics Optimizer that can decide whether to replicate transactions and when to apply them based on application semantics if such is available, the actual content of the transaction, and whether the replica is in a primary component or not. Postgres specific interceptor is the interfaces of the replication engine with the DBMS client-server protocol. Existing applications can transparently use interceptor layer to provide interface identical to the Postgres interface, while the Postgres database server sees interceptor as a regular client. The database itself does not need to be modified nor do the applications. Fig. 2 shows Postgres replication structural design.

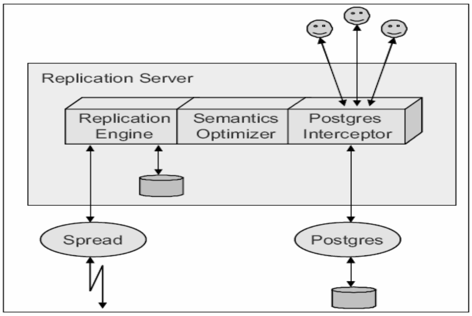


Fig. 2. Postgres replication structural design [18]

3 ROWA-MSTS Framework and Process Flow

Five phases involve in ROWA-MSTS process flow in order to preserve the replicated data consistency. This includes initiate lock; propagate lock; obtain a quorum;, update data and commit. Fig.3 shows the process flow of ROWA-MSTS in order to preserve replicate data consistency.

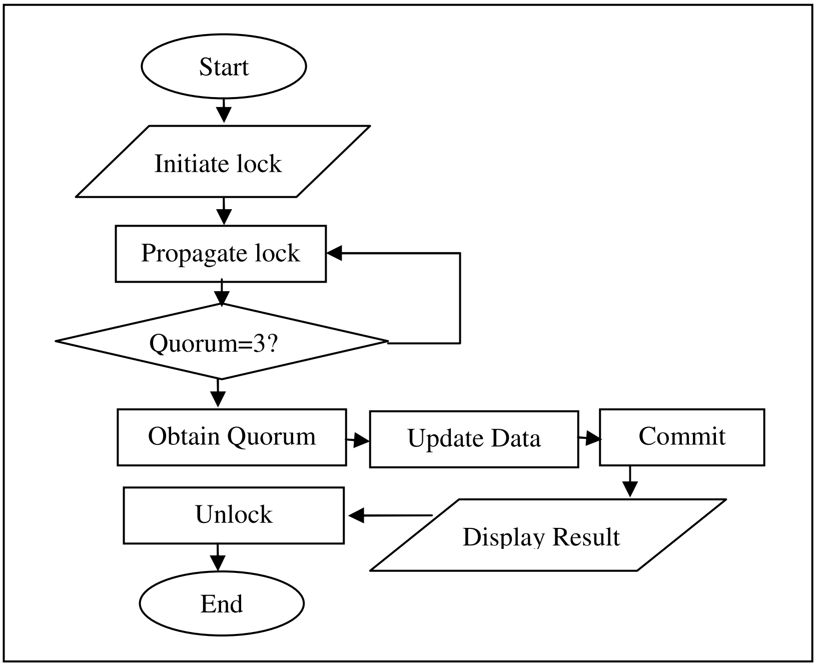


Fig. 3. Process flow of ROWA-MSTS to preserve replicated data consistency

Meanwhile, ROWA-MSTS becomes as the medium application that handle the replication consistency. Fig. 4 shows the framework of ROWA-MSTS in distributed environment.

To get intuitive ideas on how ROWA-MSTS function, let say we consider the following case. A replicated data exists namely data *x* with 3 servers in distributed environment. Each server is communicated with one another. The following notations are defined:

- PC A, PC B, PC C are the servers.
- *y is* transmitted data.
- *p, q* are the given locations.
- *Tp* and *Tq* are synchronous transaction at given locations.
- *t* is a specific time.

Let say at time *t=1*, *Tp* request to update *y=2* at server B. and *Tq* also request to update *y=3* at server C at the same time where *t=1*. The situation makes conflicts occurred since the concurrent update transactions have been invoked for the particular data *X*. If there is no concurrency control for this situation, *Tp* will update data *y*=2 at server B and *Tq* will update the value of *y*=3 at server C at the same time, *t*=1. At time *t*=2, *y*=2 at server B, *y*=3 at server C and *X*=1 at server A. Thus, at time, *t*=2 data is not at consistent state. An issue arises, how we maintain coherency and consistently transaction between all replicas in different servers at one time. In addition, we also need to make sure that synchronize replicated data can be transferred correctly.

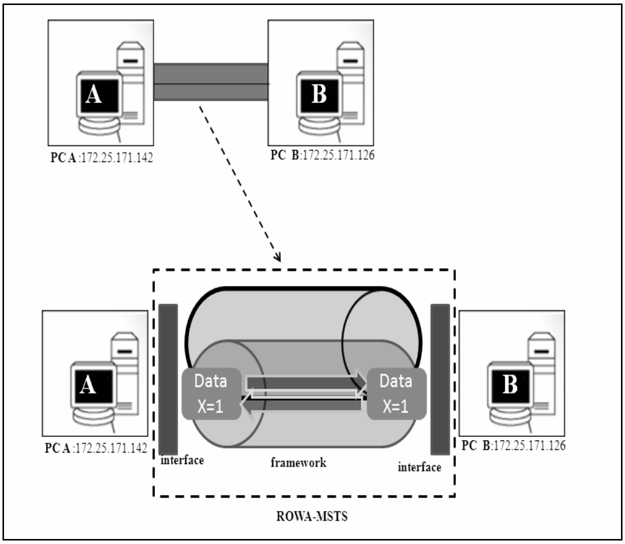


Fig. 4. The framework of ROWA-MSTS

4 Implementation and Result

In this experiment, no failures are considered during the transaction execution. The experiment aims to preserve the file consistency of during the execution. Three (3) replication servers are deployed in this implementation. Each of the servers was connected to each other via fast Ethernet router. Table 1 shows the coordination between master and neighbour coordination. Meanwhile, Table 2 shows the status lock set in ROWA-MSTS.

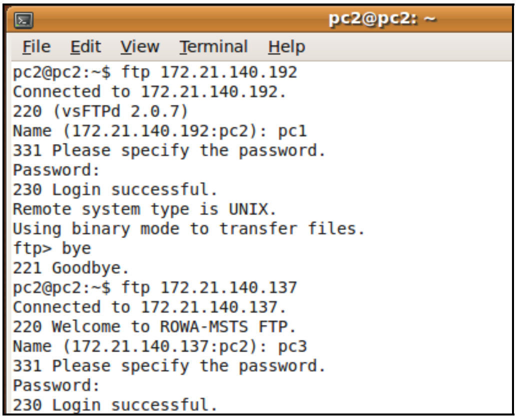
Table 1. ROWA-MSTS Master-Neighbour Coordination

Primary	**Neighbour**	
A : 172.21.140.223	B : 172.21.140.137	C: 172.21.140.192

Table 2. ROWA-MSTS Lock Status

Lock Status	Description
0	Server ready to accept any transaction from other server.
1	Server is not available for any transaction or might probably busy with other transaction.
-1	Server initiates the transaction or become the primary server.

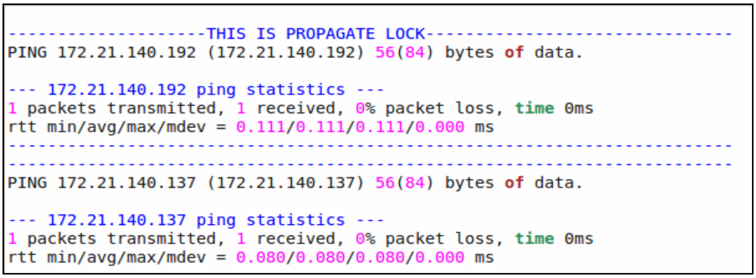
ROWA-MSTS is developed in a distributed system environment. Therefore it is necessary to gradually check the connection between all of the servers. To check the connection, the ROWA-MSTS scripting programming ping each of the connected servers. The experiment of ROWA-MSTS program was done in shell integrated with File Transfer Protocol (FTP) for the communications agent. Meanwhile, an VSFTPD is used in shell programming for receiving and sending agents. All applications for users are available to these Linux platforms. As such, the applications for users include *gedit* and *vim editor*. Before the program is executed, it is very vital to examine the FTP connection successful between all servers. Fig. 5 shows the test FTP connection successful for neighbour C and neighbour B with IP address 172.21.140.192 and IP address 172.21.140.137 respectively.



```
pc2@pc2:~$ ftp 172.21.140.192
Connected to 172.21.140.192.
220 (vsFTPd 2.0.7)
Name (172.21.140.192:pc2): pc1
331 Please specify the password.
Password:
230 Login successful.
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> bye
221 Goodbye.
pc2@pc2:~$ ftp 172.21.140.137
Connected to 172.21.140.137.
220 Welcome to ROWA-MSTS FTP.
Name (172.21.140.137:pc2): pc3
331 Please specify the password.
Password:
230 Login successful.
```

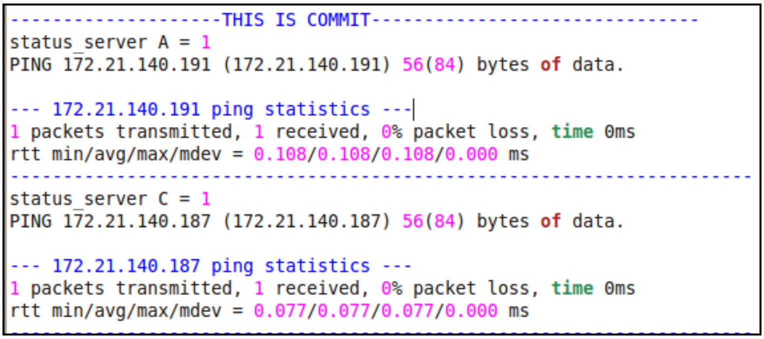
Fig. 5. FTP connection successful

The scripting has been auto configured at */etc.rc.d* so that it will be run automatically when booting process. During the execution, ROWA-MSTS generates log file to see the current status of replication job. Figure 6 and Figure 7 show the log file report of ROWA-MSTS during the propagate lock and commit data respectively.



```
--------------------THIS IS PROPAGATE LOCK------------------------------
PING 172.21.140.192 (172.21.140.192) 56(84) bytes of data.

--- 172.21.140.192 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.111/0.111/0.111/0.000 ms
-----------------------------------------------------------------------
-----------------------------------------------------------------------
PING 172.21.140.137 (172.21.140.137) 56(84) bytes of data.

--- 172.21.140.137 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.080/0.080/0.080/0.000 ms
```

Fig. 6. ROWA-MSTS during propagate lock phase



```
--------------------THIS IS COMMIT-------------------------------
status_server A = 1
PING 172.21.140.191 (172.21.140.191) 56(84) bytes of data.

--- 172.21.140.191 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.108/0.108/0.108/0.000 ms
-----------------------------------------------------------------------
status_server C = 1
PING 172.21.140.187 (172.21.140.187) 56(84) bytes of data.

--- 172.21.140.187 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 0.077/0.077/0.077/0.000 ms
```

Fig. 7. ROWA-MSTS during commit phase

Table 3. ROWA-MSTS Handle The Transaction

Replica	A	**B**	**C**
Time			
t1	Begin transaction *Initiate lock:* Write_lock(*y*) Count_write(*y*) = 1 Wait		
t2		*Propagate lock y* to server B	*Propagate lock y* to server C
t3		Lock (*y*) from A	Lock (*y*) from A
t4	*Get lock y from B* *Check quorum:* Count_write(*y*) =2 Get lock *y* from C *Check quorum:* Count_write(*y*) = 3		
t5	*Obtain quorum:* Quorum =3		
t6	Acknowledged client Update *y*		
t7	*Commit T*	*Commit T*	*Commit T*
t8	*Unlock(y)*	*Unlock(y)*	*Unlock(y)*

Consider a case where a transaction comes to update data file *y* at server A. The ROWA-MSTS implement the proposed process flow and framework during the file replication. Initiate lock occurs at Server A since invoking transaction at server A and data *y* has free lock. Thus, the write counter is increased 1. ROWA-MSTS propagates lock to server B and server C. Next, it checks either successful obtains quorum or not. Since the quorum is equivalent to three and issued from server A to get quorum, ROWA-MSTS acknowledges client that request to update data *y*. After client finished updating the data, ROWA-MSTS commits data at all servers, then unlock the data *y*. Table 3 shows the experiment result of how ROWA-MSTS handle the replication during the transaction execution.

5 Conclusion

A novel contribution of this paper is a new process flow and framework to manage replication and transaction using ROWA-MSTS. The proposed framework through ROWA-MSTS deploys real time application in distributed systems environment. Implementation shows that ROWA-MSTS able to maintain and preserved the replicated data consistency over multiple sites.

Acknowledgement

This research has been supported by University Malaysia Pahang under grant RDU080328.

References

1. Pérez, J.M., García-Carballeira, F., Carretero, J., Calderóna, A., Fernándeza, J.: Branch replication scheme: A New Model for Data Replication in Large Scale Data Grids. Future Generation Computer Systems 26(1), 12–20 (2010)
2. Gao, L., Dahlin, M., Noraziah, A., Zheng, J., Iyengar, A.: Improving Availability and Performance with Application-Specific Data Replication. IEEE Trans. Knowledge and Data Engineering 17(1), 106–200 (2005)
3. Mat Deris, M., Noraziah, A., Saman, M.Y., Noraida, A., Yuan, Y.W.: High System Availability Using Neighbour Replication on Grid. Special Issue in IEICE Transaction on Information and System Society E87-D(7) (2004)
4. Tang, M., Lee, B.S., Tang, X., Yeo, C.K.: The Impact on Data Replication on Job Scheduling Performance in the Data Grid. International Journal of Future Generation of Computer Systems (22), 254–268 (2006)
5. Noraziah, A., Fauzi, A.A.C., Sidek, R.M., Zin, N.M., Beg, A.H.: Lowest data replication storage of binary vote assignment data grid. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. Communications in Computer and Information Science, vol. 88, pp. 466–473. Springer, Heidelberg (2010)
6. Budiarto, S.N., Tsukamoto, M.: Data Management Issues in Mobile and Peer-to-Peer Environment. Data and Knowledge Engineering 41, 183–204 (2002)

7. Noraziah, A., Zin, N.M., Sidek, R.M., Klaib, M.F.J., Wahab, M.H.A.: Neighbour replica transaction failure framework in data grid. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. CCIS, vol. 88, pp. 488–495. Springer, Heidelberg (2010)
8. Bsoul, M., Al-Khasawneh, A., Abdallah, E.E., Kilani, Y.: Enhanced Fast Spread Replication Strategy for Data Grid. Journal of Network and Computer Applications 34, 575–580 (2011), doi:10.1016/j.jnca.2010.12.006.
9. Sun, X., Zheng, J., Liu, Q., Liu, Y.: Dynamic Data Replication Based on Access Cost in Distributed Systems. In: 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology. IEEE, Los Alamitos (2009), doi:10.1109/ICCIT.2009.198
10. Noraziah, A., Klaib, M.F.J., Sidek, R.M.: Failure Semantic of Neighbour Replication Grid Transaction Model. In: 2010 10th IEEE International Conference on Computer and Information Technology, pp. 668–673 (2010), doi:10.1109/CIT.2010.132
11. Thomas, R.H.: A Majority Consensus Approach to Concurrency Control for Multiple Copy Databases. ACM Transactions Database System 4(2), 180–229 (1979)
12. Noraziah, A., Mat Deris, M., Ahmed, N.A., Norhayati, R., Saman, M.Y., Zeyad, M.A.: Preserving Data Consistency through Neighbor Replication on Grid Daemon. American Journal of Applied Science, Science Publications 4(10), 748–755 (2007)
13. Holliday, J., Steinke, R., Agrawal, D., El-Abbadi, A.: Epidemic Algorithms for Replicated Databases. IEEE Trans. on Know. and Data Engineering 15(3), 1–21 (2003)
14. Stockinger, H.: Distributed Database Management Systems and The Data Grid. In: IEEE-NASA Symposium, pp. 1–12 (2001)
15. Noraziah, A., Mat Deris, M., Saman, M.Y.M., Norhayati, R., Rabiei, M., Shuhadah, W.N.W.: Managing Transaction on Grid-Neighbour Replication in Distributed System. International Journal of Computer Mathematics 86(9), 1624–1633 (2009)
16. Noraziah, A., Sidek, R.M., Klaib, M.F.J.: Development of ROWA-MSTS in Distributed System Environment. In: The 2010 International Conference on Computer Research and Development (ICCRD 2010), vol. 1(1), pp. 868–871 (2010)
17. Buretta, M.: Data Replication: Tools and Techniques for Managing Distributed Information. John Wiley, New York (1997)
18. Amir, Y.: On the Performance of Wide-Area Synchronous Database Replication, Technical Report CNDS-2002-4 (2002)

Halal Product Price Indicator Portal: Promoting Malaysia's Local Halal SMEs

Alfi Khairiansyah Machfud, Jawdat Khatib,
Ahmed Abdulaziz Haji-Ahmed, and Abdul Rahman Ahmad Dahlan

International Islamic University Malaysia,
Department of Information Systems,
Kuala Lumpur, Malaysia
alfi.khair@gmail.com, jawdat.khateeb@yahoo.com,
aboabas2004@hotmail.com, arad@iium.edu.my

Abstract. Local Halal small and medium enterprises (SMEs) play an important role in Malaysia's future economy. Currently the popularity of their products is still low compare to those of the large and multinational companies. A Halal Product Price Indicator Portal is proposed to help promote and improve the condition of the SMEs. The portal involves Malaysia Department of Islamic Development (JAKIM) and local Halal SMEs in Malaysia. The main feature of the portal is Halal products price information and comparison functionality. It is believed that the establishment of the portal will encourage people to view it, and in time, will help promote local Halal SMEs and made their products more accessible to customers and eventually contribute to national development.

Keywords: Halal Product, Price Indicator Portal, Halal SME, JAKIM.

1 Introduction

Halal is a supreme term for all Muslims across the globe. "Halal" in Arabic means permissible according to Islamic law. It is a term designated to any products or services. Halal today is a global industry. Companies all over the world have amplified their income from food exports by indulging the growing Muslim consumer market. The annual global market value for the entire Halal trade is USD2.1 trillion [1].

As stated in the Third Industrial Master Plan, Malaysia is aiming to become a global hub for the production and trade of Halal products and services. Therefore, the government encourages the establishment and the development of Halal small and medium enterprise (SME). Currently, there are approximately 2000 halal certified industries and more than 177 thousand certified products [2].

Although the Malaysian government has tried many efforts in helping the local halal SMEs, the popularity of the products is still low compared to those of big or multinational companies. Many websites have been developed to boost the popularity of Halal industry, but unfortunately only few have been built specifically to help promote the local Halal SMEs' products. This paper proposes a portal concept to help

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 254–262, 2011.

promote the local Halal SMEs. This paper also briefly discusses Halal definition, business, certification, and current issues in Malaysia. Finally, the Halal Product Price Indicator Portal model and its benefits is illustrated and explained.

2 Literature Review

2.1 Definition of Halal

In Arabic, Halal generally means permissible or authorized according to Islamic law. It refers to things or actions that will not imposed punishment to the doer [3]. A general rule in Islamic law is that everything is Halal, except if stated otherwise. Halal is paramount for every Muslim across the globe. A Muslim should be able to determine which is Halal and which is not. Nowadays, the area of Halal covers not merely food, but has extended to cosmetics, pharmaceutical, and even services such as finance. The opposite of Halal is Haram; it is defined as what is forbidden according to Islamic law. However, the term non-Halal is often used in preference of Haram. In Malaysia for example, non-Muslims restaurants and food are signed non-Halal [4].

2.2 Halal Business

Halal today has been included in the area of business and trade. It has even become a new benchmark for safety and quality assurance [5]. Certified Halal products are generally acceptable by both Muslim and non-Muslim consumers [6]. This acceptance is due to the wholesomeness concept of Halal, which covers not only the Islamic law requirements, but also the requirements for good food, in terms of hygiene, sanitation and safety factors [5].

The global Halal industry has become enormous. According to HDC [1] The annual global market value for the entire Halal trade reach USD2.1 trillion. In UK alone, the retail sales of Halal meat grasp USD90 million. In Asia, with a Muslim population of approximately 1 billion, the prospect is even bigger. Knowing the potential of Halal industry, the Malaysian government focused on making the country an international Halal hub. In the Ninth Malaysia Plan, Malaysia is aiming to be a stop center for Halal certification worldwide and positioned as the knowledge center for trade and investment promotion of Halal products. For that reason, the government established the Halal Industry Development Corporation (HDC) and organize annual international event such as the Malaysia International Halal Showcase (MIHAS) and World Halal Forum (WHF) [2]. SMEs involved in activities such as food processing, pharmaceuticals, and other consumables are encouraged to obtain Halal certification to leverage the country's competitive edge in becoming the Halal hub. The government even provides assistance to companies in a form of grant as much as RM150,000 per company for the development and promotion of Halal products, and another RM250,000 for productivity and quality improvement (Third Industrial Master Plan). As a result, the number of Halal-certified enterprises started to grow and in

2010, it grew significantly by 20 percent, from 1,399 companies in 2009 to 1,679 companies. 65 percent of these industries are from food clusters [7].

2.3 Halal Certification in Malaysia and Its Benefits

Unlike other countries around the world that have private Islamic organizations issuing Halal certificates, Malaysia's Halal certificate is issued by the government, particularly by the Malaysia Department of Islamic Development (JAKIM). Having the government controlling and monitoring the Halal certificate made Malaysia's Halal brand much stronger than others [6]. According to Yusoff [8] Halal food certification refers to *"the examination of food processes in its preparation, slaughtering, cleaning, processing, handling, disinfecting, storing, transportation and management practices."* The Halal concept, specifically in terms of food, should apply to all stages of its processes. In short, "from farm to table."

The benefits of Halal certification are clear: (1) It provides confidence to consumer in consuming the products, (2) it can be used as a marketing tool and provides the competitive advantage for the manufacturers, (3) it indicates that not only the product satisfy Islamic law requirements, but also adheres to stringent hygiene and sanitation practices, and (4) it also provides a mechanism for the authority to audit and monitor Halal food [8]. A study of consumer behavior in Malaysia by Danesh et.al. [6] shows that both Muslims and non-Muslims accept and consume Halal products, and they also satisfied with the quality. For manufacturers, the impact is more evident. In 2007, since the United Arab Emirates recognition of MUIS' (Islamic Religious Council of Singapore) Halal certification system, the country's food exports to UAE rocketed by 67% in just one year [9].

2.4 Web Portal

A web portal is a term for a website that derived information from various resources and presented in an integrated form. It can accommodate large audiences, which can be translated to a large number of advertising viewers. Murray [10] expresses that information portal has the ability to organize large context based information and connects people with it. According to White [11] The basic architecture of an information portal mainly comprise of the business information directory, the search engine, the metadata crawler, the publishing and subscription facilities, and the import and export interfaces, all integrated in a web server.

According to Dias [12], portals can be classified according to their type of environment (public or corporate) and their functions (decision support and/or collaborative processing). The portal discussed in this paper should fit the category of public in terms of its environment, and collaborative processing in terms of its functions.

2.5 Price Indicator Information System

Halal product price indicator in this paper is defined as a range of prices of Halal certified products produced by local SMEs in Malaysia. The objective of the price indicator is to make price information more widely available for consumers as well as businesses in order to improve the campaign of Halal products.

The price indicator information system in this paper refers to a specific feature that presents comprehensive information about Malaysia's local Halal product prices. The idea was inspired by the price indicator information system found in the website of Ministry of Trade and Industry of Saudi Arabia (http://www.cpi.mci.gov.sa) and also in the website of Ministry of Business and Trade of Qatar (http://market.mbt.gov.qa/). According to Al-Abosh [13], in early 2010, The Ministry of Trade and Industry of Saudi Arabia launched the consumer price indicator information system. The system provides consumers information on various types of products and commodities, and also the price levels and comparison between products. The prices are updated on a daily basis. With this system consumers are able to choose suitable product that fits their level of income and they can also decide from which outlet or store they want to buy. In addition, the system also provides information to researchers and analyst on the movement of commodity prices through a series of time. Similar system can be found in Qatar's Ministry of Business and Trade website, but with additional English language and more variety of goods. Everyday the people of these two countries viewed and use this portal as they get many benefits from it.

Regarding the notion to promote local Halal products in Malaysia, by having the price indicator feature similar to the ones in Saudi and Qatar above, but built specific for local Halal products, customers will be encouraged to view the portal on a regular basis according to their needs. Therefore, the popularity of the local Halal products in the market will be increased gradually over time. It is believed that the more popular the product, the more the demand by the market, which will result in the growth of the local Halal SMEs.

3 Current Issues

Malaysia is a potential growth place for Halal SMEs as the government fully supports the establishment and development of Halal product and industry. Many efforts have been done to accomplish the country's goal on becoming the world's Halal hub. From policies written in the third industrial master plan, grants and annual events, to the establishment of relevant organization such as Halal Industry Development Corporation (HDC). However, as stated by Sobian [14], the awareness of the consumers towards local or domestic product is still low. Lack of information on most of the Halal products in the market is the main factor that contributes to this situation. Another issue is that even though 85% of the Halal certified companies are SMEs and 68% of certified Halal products also come from SMEs [2], most of the Halal products well-known in the market are those from the big or multinational companies. This situation maybe due to their large capital and marketing experience in the business industry.

In terms of promotion, currently there are several websites that promote Halal SMEs in Malaysia. However, according to our observation, these websites merely provide general information, news and company profile. There is no specific website that provide local Halal products price information and at the same time, distinguishes between local or Muslim companies and those who are not. In addition, most of these websites also promote Halal products and companies originate from outside Malaysia.

Based on the above arguments, there is a need to leverage the current condition of the local Halal SMEs in Malaysia and promote their products. One method that can be used is by establishing a Halal portal with price indicator information system specifically built for local Halal products and companies.

4 Halal Product Price Indicator Portal Model

The proposed portal will incorporate government agency such as the Malaysia Department of Islamic Development (JAKIM) and local Halal SMEs, particularly manufacturers and shops. Figure 1 shows the relationships between local Halal SMEs, JAKIM, and potential customers through the portal.

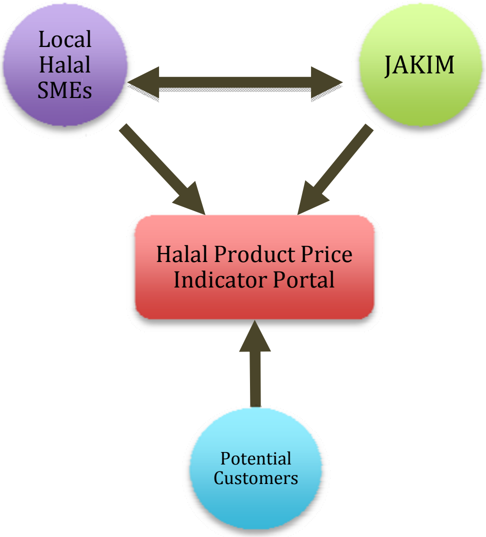


Fig. 1. Halal Product Price Indicator Portal Model

JAKIM is the sole authority for Halal certification in Malaysia. Since it has the database of every Halal certified manufacturers and products, it is the most appropriate organization to manage the portal. Local Halal SMEs refers to every certified manufacturer and shop, which is owned by local or Muslim citizens of Malaysia. These companies interact with JAKIM when they apply for Halal certification. The portal will be the center point where the local Halal SMEs, JAKIM and potential customers can benefit from. Figure 2 describes the relationships more detail between the portal and JAKIM, local Halal SMEs and potential customers. We consider it is more appropriate to categorize the local Halal SMEs to manufacturers and stores, as they may sell the same product but with different prices. Generally, stores will be selling Halal products with retail price while manufacturers with wholesale price.

As can be seen from Figure 2, manufacturers and stores interact with the system by providing registration details, their products and its prices. In return, the system will provide them with subscription approval. The companies can also update their profiles and product prices on a regular basis. The system provides potential customer with local Halal SMEs' profiles including their products and prices. JAKIM has the role of giving the portal Halal product and manufacturers details together with their ownership status, as it will differentiate between local or non-local Halal SMEs.

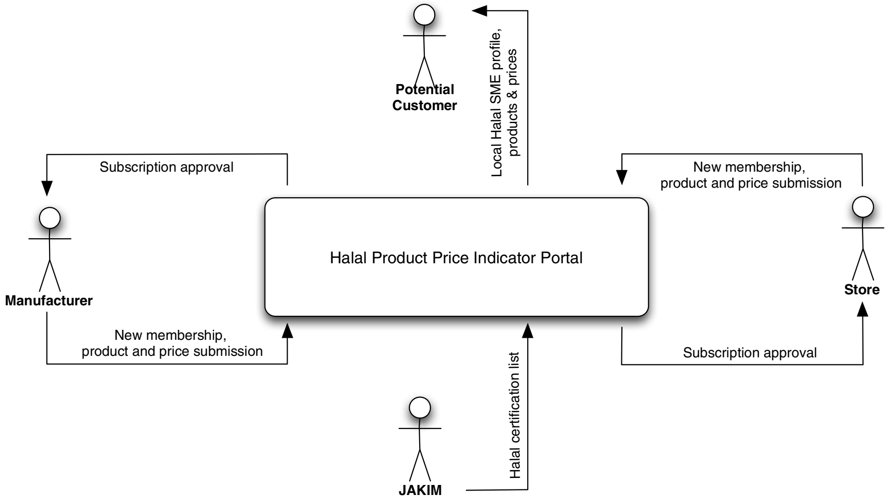

Potential Customer
Local Halal SME profile, products & prices
Subscription approval
New membership, product and price submission
Manufacturer
Halal Product Price Indicator Portal
Store
New membership, product and price submission
Halal certification list
Subscription approval
JAKIM

Fig. 2. Relationships between the portal and its components

Figure 3 illustrates the functionality of the portal by defining its users, functions and relationship between them. Manufacturers and shops have a general term local Halal SME. A local Halal SME can subscribe for a membership where after being approved, has a choice to update its profile. Product prices can also be updated. The local Halal SME will then have to update the prices on a regular basis. A potential customer can view manufacturers or shops profiles, view their products' prices, and also able to compare prices between products in the same category. He or she can also view product price history or movement of a certain product. JAKIM will upload the Halal certification database and it will also include the SME ownership profile (e.g. whether it is local or not). This action will filter the local Halal SMEs.

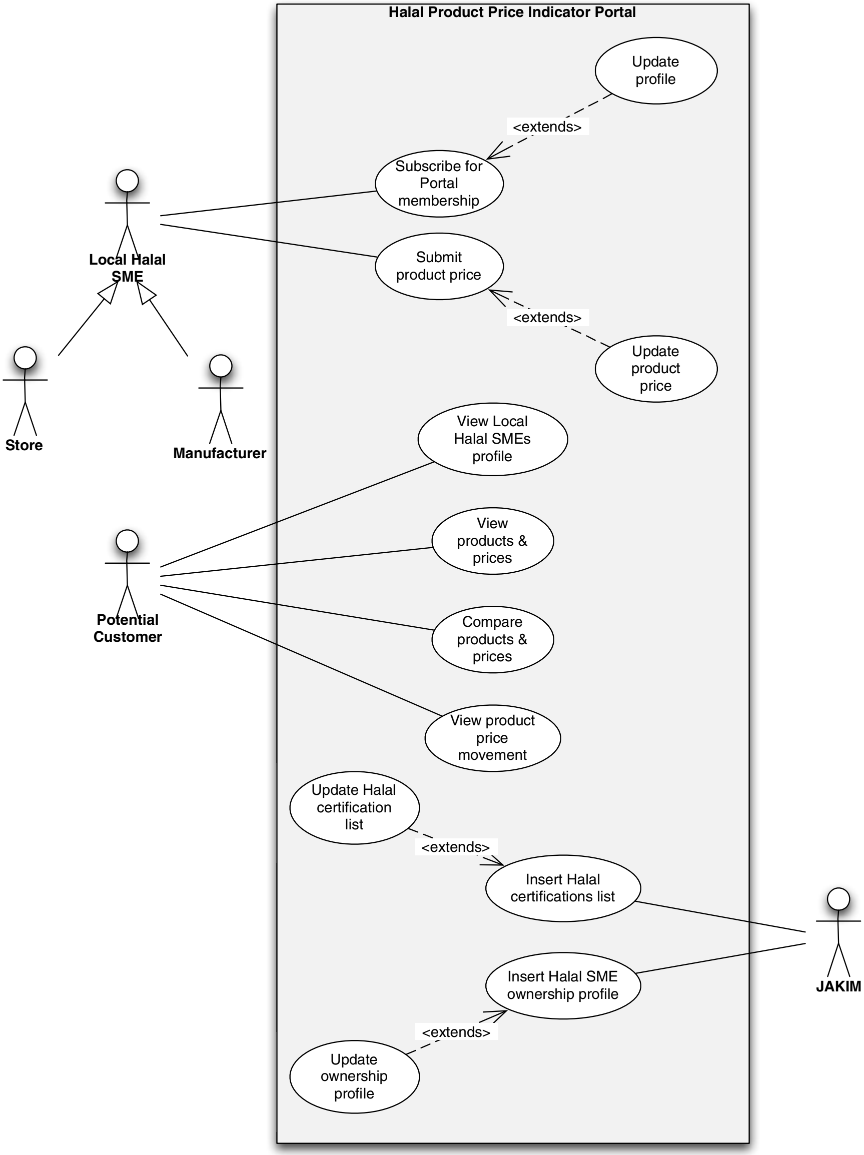

Halal Product Price Indicator Portal
Update profile
<extends>
Subscribe for Portal membership
Submit product price
<extends>
Update product price
View Local Halal SMEs profile
View products & prices
Compare products & prices
View product price movement
Update Halal certification list
<extends>
Insert Halal certifications list
Insert Halal SME ownership profile
<extends>
Update ownership profile
Local Halal SME
Store
Manufacturer
Potential Customer
JAKIM

Fig. 3. Portal functionality and its components

5 Benefits of the Portal

The Halal Product Price Indicator Portal will benefit all stakeholders involved. The local SMEs will have a new market place to present their products. The portal will act as a media of promotion for their products. This will help them penetrate new markets they cannot reach within their current capabilities. Customers are the direct beneficiaries of the portal as they can browse through a variety of local Halal products made by manufacturers from a reliable source. At the same time, they can compare between them in terms of prices and specifications before deciding to purchase from a certain physical store or even directly to the manufacturer. A feature of product price history and prediction in the portal will tell customers the direction of prices depending on their movement. As a result, customers can decide to purchase now or delay the purchase according to product price prediction. JAKIM as the one who maintain and manage the portal will benefit indirectly from it by promoting themselves as trusted major regulators in the Halal product market.

In terms of social and economic value, the establishment of the portal promotes price transparency and fairness, helping local Halal SMEs to find their roots in a very competitive market. The portal will also allows middle-class customers to compare between prices in order to decide what and where they should purchase the Halal product suitable for them.

6 Challenges in Implementation

Several challenges may arise in implementing the Halal Product Price Indicator Portal. First, identifying local and Muslim owned manufacturers and suppliers can be quite difficult. Some manufacturers are built as a joint stock company, which means it has more than one owner (e.g. part of the company is owned by a non-local or a non-Muslim); making the ownership status of the manufacturer ambiguous. The second challenge is to enforce the local Halal SMEs to update their product prices on a regular basis, as they may not comply to update their products. In addition, there is also a chance that the companies submit false prices. Thus, an appropriate policy must be constructed and actively enforced by JAKIM to establish and ensure customers' trust on the portal.

7 Concluding Remarks

The proposed Halal Product Price Indicator Portal concept can provide a place for local Halal SMEs in Malaysia to promote their products in the market. It is believed that the involvement of relevant government institution will add value to the portal and nurture trust from customers. The future concept of the portal can be extended to provide more features and functionalities such as e-commerce or marketplace where customers or businesses can interact and purchase products directly from the portal. However, more research on this issue is definitely needed.

References

1. Halal Industry Development Corporation. Halal Market Information, http://www.hdcglobal.com/publisher/bhi_market_information
2. Halal Industry Development Corporation. Halal Industry Statistics, http://www.hdcglobal.com/publisher/alias/bhi_industry_statistics?dt.driverAction=RENDER&pc.portletMode=view&pc.windowState=normal&pc.portletId=hadirGraph-web.hadirWebPortlet

3. The Malaysia Government Open Portal. Halal Definition, `http://www.malaysia.gov.my`
4. Wilson, J., Liu, J.: Shaping the Halal into a brand? Journal of Islamic Marketing 1 (2010)
5. Bistari, M.Z.: Standards & Quality News, p. 1 (July-August 2004)
6. Danesh, S.M., Chavosh, A., Nahavandi, M.: Comparative Analysis of the Muslims' and Non-Muslims' satisfaction with Halal products (2010)
7. Halal Journal Team. OIC Eyes The USD580 Billion Global Halal Market. Halal Journal Team (2008), `http://www.halaljournal.com/article/1409/-oic-eyes-the-usd580-billion-global-halal-market`
8. Yusoff, H.: Halal Certification Scheme: Standards & Quality News, pp. 4–5 (July-August 2004)
9. International Enterprise Singapore. Fast-growing food exports provided a boost to total food trade with UAE by more than 60% (2007), `http://www.iesingapore.gov.sg`
10. Murray, G.: The portal is the desktop. Intraspect, May-June (1999); Dias C.: Corporate portals: a literature review of a new concept in Information Management. International Journal of Information Management 21, 269–287 (2001)
11. White, C.: Enterprise information portal requirements. Decision processing brief. Morgan Hill (1999); Dias C.: Corporate portals: a literature review of a new concept in Information Management. International Journal of Information Management 21, 269–287 (2001)
12. Dias, C.: Corporate portals: a literature review of a new concept in Information Management. International Journal of Information Management 21, 269–287 (2001)
13. Al-Abosh, K.: Saudi Arabia launches index of consumer prices in their domestic markets: Arabia Business (2010), `http://www.arabianbusiness.com/arabic/580084`
14. Sobian, A.: Perbaiki Sistem Pemasaran Produk Halal. In: Nooh, M.N., Nawai, N., Dali, N.R.S., Mohammad, H. (eds.) Proceedings of the 1st Entrepreneurship & Management International Conference Certification: What the SME Producers Should Know, Kangar Perlis, December 5-7 (2007)

A Pattern for Structuring the Information System Architecture

-Introducing an EA Framework for Organizing Tasks

Shokoofeh Ketabchi, Navid Karimi Sani, and Kecheng Liu

Informatics Research Center (IRC), University of Reading, PO Box 241,
Whiteknights, Reading, RG6 6WB, UK
{s.ketabchi,n.karimisani,k.liu}@reading.ac.uk

Abstract. Enterprise Architecture (EA) is an approach which aims to align IT and business strategy to help organizations invest wisely in IT and make the most of their current IT facilities. In fact, EA is a conceptual framework which defines the structure of an enterprise, its components and their relationships. It is considered to be a massive and also complicated job which if proper methods and approaches are not used, a huge amount of chaos and overheads will be caused. This paper aims to introduce a structure for organizing EA tasks. This structure provides guidelines on how to organize and document business services. Therefore, a more organized EA process will be carried out and both time and effort are saved. A case study is used to elaborate and explain the structure accompanied by the evaluation and discussion.

Keywords: Enterprise Architecture, Service Oriented Architecture, EA, SOA, BAITS, Information System Architecture.

1 Introduction

Organizations face an increased level of competitiveness nowadays. IT plays an important role to help organizations stay competitive and moving forward. The alignment of IT with business is one of the main issues. Enterprise Architecture (EA) helps to align Business strategy with IT. As mentioned by Kaisler *et al.* (2005) "*Organizations need to have clear, but concise, strategic plans for business and IT. The business strategic plan becomes the driver for the EA*" [12]. Therefore, the main purpose of EA is to create a unique IT environment within an enterprise. EA is sometimes accompanied by the term Service –Oriented Architecture (SOA). Bell (2007), one of the leading Enterprise architects, defines EA as a city planner, "one who oversees how the entire landscape comes together" [2]. In fact, when entire city is considered and how different sections fit into the city. Conversely, he believes SOA focuses on **the** "delivery of city services and facilitating communication within the city" [2].

There are a vast majority of frameworks and guidelines on how to do EA and how to align business and IT. However, there is still not a common solution. Frameworks are short of architectural parts definition. Moreover, it is quite difficult to

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 263–273, 2011.

communicate the outcome of EA to the whole organization since connections are not well understood or documented [11].

BAITS (Business-Aligned IT Strategy) is a set of well-defined methods and tools towards IT-business alignment. BAITS helps organization to achieve a holistic view of their organizations and different sections. All core services, their relationships and stakeholders' interactions with those services will be clarified.

As a result, a map will be produced and each individual can point out the section they are working on. Thus, it will facilitate coordination. This approach helps to align IT and business more efficient and effective.

This paper provides an introduction on BAITS approach. The main objective is to show how BAITS is different from other EA frameworks and how it adds value to organizations. Section two provides some backgrounds on EA and SOA along with related work and the concept of business-IT alignment. Third section defines BAITS, its structure and phases. Section four shows a case study with the results of BAITS being applied in a particular domain; followed by section five which provides validation and discussion on the outcome of BAITS application. Finally, a brief conclusion of the whole paper is provided in section six.

2 EA and SOA

Enterprise Architecture (EA) can be defined as a representation of all enterprise components with their interactions and relationships. "*It is a complete expression of the enterprise*" [24]. Rood (1994) defines EA as the representation of components and their interactions within an enterprise. He believes that "*EA is a conceptual framework that describes how an enterprise is constructed by defining its primary components and the relationships among these components"* [23]. EA helps to have centralized and consistent information within an enterprise. Thus, duplication and inconsistencies will be reduced. This brings some benefits for the enterprise. First, this information is of high quality and available whenever and wherever needed to make better decisions. Second, it helps to invest wisely in IT and, in fact; it improves *Return on Investment (ROI) for future IT implementation*" [15]. Therefore, EA can reduce cost and keep enterprises competitive. The EA deliverables includes models, graphics, representation and description/documentation for future roadmap and IT strategy.

2.1 Service Oriented Architecture (SOA)

There is no commonly agreed definition for SOA. IBM defines it as "*an enterprise-scale IT architecture for linking resources on demand. The primary structuring element for SOA applications is a service as opposed to subsystems, systems, or components*" [10]. Therefore, SOA simply identifies services, stakeholders and their interactions by considering whole enterprise. Jones (2006) introduces some guidelines to carry out SOA [11]. A brief explanation is provided below; each phase answers a particular question:

1. **What** (What does the company do): the services which a particular organization provides will be determined (Project Scope).
2. **Who** (Who does it): all stakeholders who are interacting with services are identified.
3. **Why**: the relationships and interactions within services and between services and stakeholders are identified.
4. **How**: how a particular service operates inside and organization is determined. For this purpose, a brief description may be written or separate diagram (could be Use-Case, Activity, Process or Sequence Diagram according to UML guidelines) may be drawn.

2.2 Related Work

Many frameworks have been developed to structure the EA implementation. They provide guidelines for and list of things that must be captured, analyzed and implemented. Zachman framework (1987) is the earliest framework. John Zachman first introduced this framework in 1987 [31] and, then, with Sowa collaboration extended it in 1992 [26]. It is well defined and considered to be "*the most referenced framework that is also a basis for evaluating, establishing and customizing other enterprise architecture frameworks, methods and tools*" [8]. Zachman introduced a matrix and guides through generating each cell to do the EA job. His website provides a practical foundation through EA implementation [30] as well as series of his workshops, seminars and conferences. According to Schekkerman, 25 percent of organizations use Zachman framework [25].

Another framework is called EAP (Enterprise Architecture Planning) which was first published by Steven H. Spewak in 1990s. He was a professional practitioner in the area of System Architecture and developed EAP as "*the process of defining architectures for the use of information in support of the business and the plan for implementing those architectures*" [27]. EAP adopts a *business data-driven* approach which intends to make sure that Information Systems are of high quality. For this purpose, it emphasizes on a developing stable business model, defining dependencies before implementing system and ordering implementation activities based on the data dependencies. The other well-known framework is, TOGAF (The Open Group Architecture Framework) has been developed by the Open Group in mid-1990. The latest version, TOGAF 9.0, was released on February 2009 [29]. The main emphasis is to produce Business, Application, Data and Technology architecture through a well defined circular structure.

Apart from well-known framework, many organizations use their own-developed EA framework to align business and IT. Kamogawa and Okada introduce a framework for EA which enhances adaptability and stability [13]. However, there is no proof that their work would increase the adaptability by hundred percent. Furthermore, Capgemini Company developed an Integrated Architecture framework (IAF) [5]. According to a survey in 2005 [25], three percent of the organizations used their framework for implementing EA. This framework provides a unique approach throughout the architecture design and also deployed Oracle E-Business suite. In addition, other frameworks such as APS (Application Portfolio Strategy) have been developed in this company which helps to align business and IT [6] [7]. This framework is further worked out with collaboration of a research group in the University of Reading and a new framework called BAITS (Business-Aligned IT Strategy) has been developed.

3 The Alignment of Business and IT Architectures

The critical component of every organization is an IT strategy which drives and facilitates business objectives. IT and business alignment, in fact, is an opportunity to obtain greater value from current and future IT investments [28]. ”*The challenge of aligning Information Technology (IT) to business has often been cited as a key issue by IT executives”* [14]. Some papers suggest different instrument to measure the alignment between IT and business. For instance, Khaiata and Zualkernan (2009) present a tool to identify maturity alignment and existing gaps [14]. This instrument “*is based on Luftman's "Strategy Alignment Maturity Model" (SAMM)* [19]; *it directly encodes all attributes of SAMM alignment areas using a unidimensional framework”* [14].Different documents are talking about different aspect of IT to be aligned with business; however, we believe if IT architecture is aligned, other parts will be aligned automatically or by a little effort.

The BAITS approach helps to understand where and identify where business value created by IT. First, all core business services and their relationships inside an organization will be determined. Then, stakeholders' interactions with them, business norms and rules will be identified. The same knowledge will be acquired about IT estate, applications and infrastructure. Afterwards, IT strategy and roadmap will be produced to provide insights into future business vision for the organization (figure 1) [3].

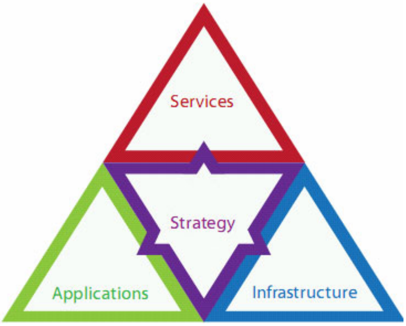


Fig. 1. Key concepts of BAITS [6]

Quite a few factors hinder the alignment of IT and business. First, there are few experts in this field that can be trusted. Second, organizational culture may hold back this process. Gregor et al. (2007) considers the social aspects of alignment [9], such as management support [22], corporate planning styles [19], and the communication of business plans to stakeholder groups [4]. Nonetheless, using a well-defined framework may solve these problems [1].

4 BAITS

4.1 BAITS in General

BAITS is an Enterprise Architecture framework which is developed by the CEAR (Capgemini Enterprise Architecture Research) group; shared group between the Informatics Research Center (IRC) in the University of Reading and the Capgemini Company. It consists of a number of analysis and modeling techniques which help to align business needs with IT efficiencies and investments. BAITS is benefited from other successful framework and methods; such as TOGAF (the Open Group Architecture Framework) and Organizational Semiotics concept and principles.

The BAITS architecture is shown in figure 2. It is composed of 4 main stages; Business Service Architecture, Portfolio Value Analysis, Portfolio Investment Analysis and Transformation Roadmap. Each stage presents several techniques to achieve corresponding goals.

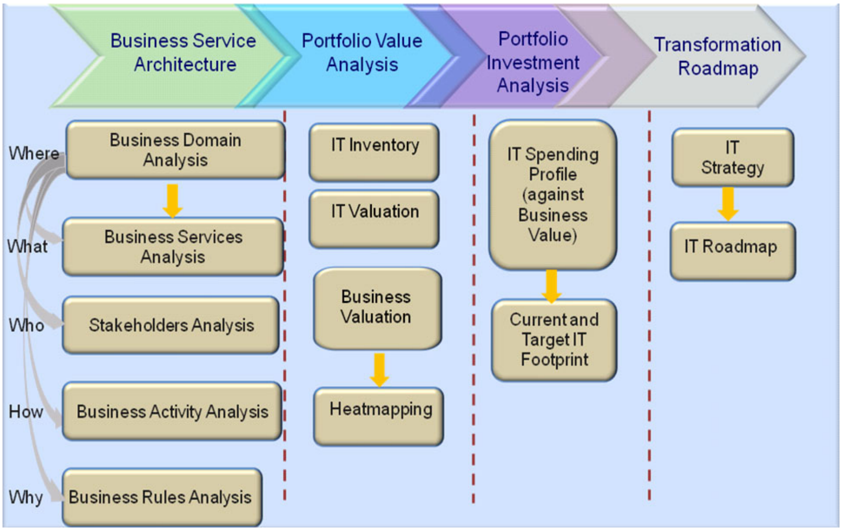


Fig. 2. BAITS Architecture [17]

The 'business service architecture' phase provides the foundation for other stages. In fact, the higher quality output from this phase will improve other phases which are more focused on qualitative and quantitative (monetary) value of services. In other words, as in any EA framework, apart from identifying service, there should be valuation and analysis about those services as is in other phases, explained in [20][21]. Explaining all stages is beyond this paper; therefore, we focus on the first phase and represent how it provides effective input for other phases in terms of quality and efficiency.

4.2 Business Service Architecture

The architecture of all services in an organization will be portrayed by accomplishing five main analysis efforts. Each analysis in this stage makes an effort to investigate a particular aspect of the system.

- Business Domain Analysis (BDA)

The main purpose is to determine the goals, mission, market condition, business processes, structure of a particular organization and also related stakeholders. It provides a holistic view of the organization and verifies the scope which analysts should work on. It is of a great importance since further analysis is based on the knowledge gained through this stage. Therefore, to align business and IT successfully, a complete knowledge of the business and its related issues must be acquired [17].

- Business Service Analysis (BSA)

It is done as the extended analysis of BDA. All core and support services within an organization will be identified, refined and combined. During this stage the question “what the company does” will be illuminated and related models will be drawn to clarify it.

- Stakeholder Analysis (SA)

All the main groups of stakeholders will be determined along with their roles, responsibilities and their impact on the organization. This phase attempts to answer the “who” question; “who does it”. A comprehensive analysis will be done to determine all relevant stakeholders according to the stakeholder onion model [18] from Organizational Semiotics approach [17].

- Business Activity Analysis (BAA)

All activities, functions, and processes which realize a particular business service and also those which operate across business services will be identified.

- Business Rule Analysis (BRA)

All norms and rules which govern the whole business operations of an organization will be specified. The agent (subject) along with conditions and impediments in which a norm should or should not be realized will also be determined.

Different question pronouns will be answered for this purpose. The main focal point of this section is the ‘Business Service Analysis’; however, it embraces the other two analysis (‘Stakeholder Analysis’ and ‘Business Activity Analysis’). In fact, four main question pronoun will be asked and answered; ‘what’, ’who’, ’why’, ’how’.

- What: determines all business services
- Who: identifies all related stakeholders
- Why: finds out about the relations and connection between different service and their stakeholders as well.
- How: defines all processes and activities which implement a particular service and also organize services.

The outcomes of this section are input for other phases. Following section shows how these principles are applied in practice and what the outputs which act as inputs to other phases are.

5 Case Study (Part of BAITS to Help with Alignment)

The case study is a port-operator company which manages all exports and imports regarding to a particular port. This company has many services; such as Cargo Holding Equipments, Container Operation, General Cargo Operation, Sale, Finance and etc. In this paper, the high level architecture of the company is presented (Figure 3).

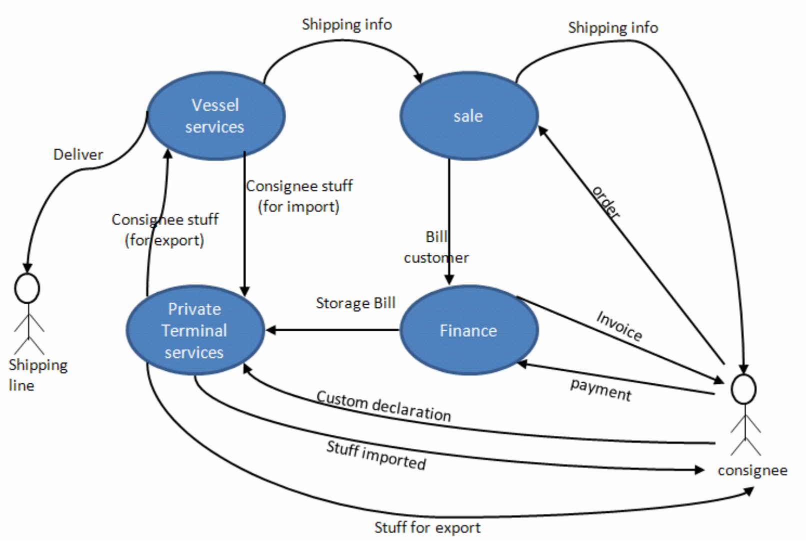


Fig. 3. Enterprise Level 0 SOA

Terminology: a vessel is a ship and consignee is a legal entity who owns the goods (stuff).

Table 1. Structure for organizing Business Service Architecture

Business Service		
Service name: Private Terminal Services	**ID**:002	**Type**: core service
	Business Value:	**Date**:
Description: all stuff should be stored here after imported or before export.		
Stakeholders: consignee		
Service Capabilities: storing stuff, issuing storage bill, verifying custom declaration of stuff.		
Process/activity Models: Private Terminal Service - activity diagram		
Relationships: Enterprise Level 0 SOA		
IT Applications: Private Terminal system	**IT Application Value to Business Service**:	
Recommendations:		

Table 1 represents all information about the ‘private terminal service’. Each field in the table1 provides a set of related information, which is gathered from different analysis and modeling techniques all across the organisation, for its related audience. Each field in table1 is explained here.

- Service name: The name of a business service.
- ID: The numbering system is used for each service.
- Type: The type of the service can be “core business service” or “support service”.
- Business Value: A business value represents a level of the strategic role of this business service to the business operation from the stakeholders view points and is calculated based on stakeholders’ feedback.
- Date: when service is documented to keep track of changes to the business service.
- Description: A brief description of what this business service is capable of doing in the business domain.
- Stakeholders: A list of stakeholders who have the input to and/or output from the business service.
- Service Capabilities: A detailed description of the service capabilities.
- Process/activity models: The models to show how the business service operates. It could be process models or activity diagram (based on UML notations [8]) to clarify the set of processes and activities that collaborate to a particular service. For example, the activity diagram in figure4 shows the series of activities needed to carry out the when a container is loaded and needs to exit from Private.

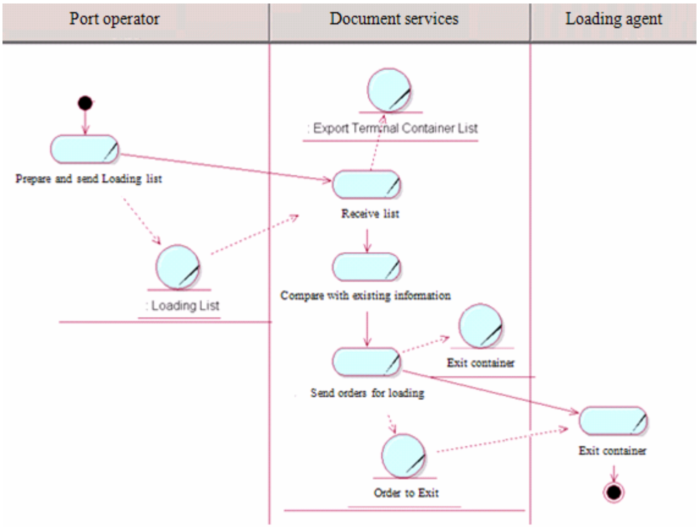


Fig. 4. loading and send out container from private Terminal Service – activity diagram

- Relationships: interactions with other business services
- IT Applications: A list of all IT applications which support the service.
- IT Application Value to the Business Service: The importance of an IT application to the business service is assessed some valuation technique.

- Recommendations: It represents the decisions having been made over the service, recommendations for future strategy decisions, etc. For example, if a business service has a high business value but a relatively low IT application value, it may indicate that the IT investment in this service should be re-considered.

6 Evaluation and Discussion

Enterprise architecture considered being complicated enough even if everything is done perfectly; however, lack of good documentation and referencing methods makes it more complicated specially in large scale projects. Service oriented architecture (SOA) is one of the successful methods for enterprise architecture. Nonetheless, like most other methods, it does not suggest any methodical approach for documentation and cross-referencing system. Lack of a documentation system in SOA can lead to chaos especially when the results (service models, process Models and etc.) are archived for future use.

BAITS as an EA framework seek to enhance SOA by suggesting patterns and approaches to document services, their process models and the relation between sub-services and super-services. This structured documentation of services can be reused more easily even by another team of analysts, without spending too much time on finding relations between documents and their process models. In other words, structured documenting improves the overall performance of BAITS as an EA framework. This feature is missing in almost all other EA frameworks.

Moreover, the BAITS framework offers a simple and familiar set of understandable and applicable techniques and methods to reduce the complexity and enhance the overall affectivity of EA projects. The first phase as described in this paper follows a simple and precise structure to document services which are identified as a result of applying these set of techniques and methods. For example, the ‘Relationships’ field in table1 shows the relationships between that particular service with other services. This information is scattered across the enterprise. Applying the BAITS framework helps to gather all this information in one place to reduce further effort for understanding basic information about a particular service, including its capabilities and relationships with other services. This is crucial in any EA framework and BAITS leverage the EA potential and competency and deliver a more effective project and outcome as a result.

7 Conclusion

This paper describes the concept of enterprise architecture (EA) and service oriented architecture (SOA) in addition to methods and approaches developed for this purpose so far. A customized method for EA, BAITS, is also introduced and explained. BAITS is a framework which attempts to guide through EA process and help with analysis and also documentation. The main focus is on introducing a structure for documenting services so that all information about a service is collected in one place and, therefore, there is an integrated view of a service with all of its sub-services, processes, stakeholders, IT systems and related decisions and recommendation in one place. This has advantages both during development and also afterwards during maintenance either by the same development team or completely different one.

Future work includes further development of the structure and providing guidelines on how to calculate values field. A more formalized approach for stakeholder analysis will also be desirable.

Acknowledgments. The authors would like to thank Dr. Sam Chong for his valuable inputs to the project and strong collaboration with the project team.

References

1. Baker, D.C., Janiszewski, M.: 7 Essential Elements of EA-Follow this roadmap to engage your business, manage complexity, and govern the implementation of your architecture by. Diamond- management and technology consultants (2005)
2. Bell, J.: Enterprise Architecture and SOA (2007), http://blogs.zdnet.com/service-oriented/?p=894
3. Business Aligned IT Strategy (BAITS) brochure, Capgemini UK (2009)
4. Calhoun, K.J., Lederer, A.L.: From strategic business planning to strategic information systems planning: the missing link. J. Information Technology Management 1(1), 1–6 (1990)
5. Capgemini: Insight and resources. Application Portfolio Strategy (2010), http://www.capgemini.com/insights-and-resources/by-publication/application_portfolio_strategy/
6. Capgemini: Integrated Architecture framework (IAF), http://www.capgemini.com/iaf
7. Capgemini: Service Oriented Architecture (2006), http://www.capgemini.com/services-and-solutions/outsourcing/application-outsourcing/solutions/application-portfolio-strategy/
8. Fatolahi, A., Shams, F.: An Investigation into Applying UML to the Zachman Framework. J. Information systems Frontier 8(2), 133–143 (2006)
9. Gregor, S., Hart, D., Martin, N.: Enterprise Architectures: Enablers of Business strategy and IS/IT Alignment in Government. J. Information Technology & People 20(2), 96–120 (2007)
10. Ibrahim, M., Long, G.: Service-Oriented Architecture and Enterprise Architecture, Part 1: A framework for understanding how SOA and Enterprise Architecture work together. IBM website (2007), http://www.ibm.com/developerworks/webservices/library/ws-soa-enterprise1/
11. Jones, S.: Enterprise SOA Adoption Strategies: Using SOA to Deliver IT to the Business. InfoQ. USA (2006)
12. Kaisler, S.H., Armour, F., Valivullah, M.: Enterprise Architecting: Critical Problems. In: The 38th Hawaii International Conference on System Sciences, HISS 2005 (2005)
13. Kamogawa, T., Okada, H.: A Framework for Enterprise Architecture Effectiveness. In: The International IEEE Conference on Services Systems and Services Management, ICSSSM 2005 (2005)
14. Khaiata, M., Zualkernan, I.A.: A Simple Instrument to Measure IT-Business Alignment Maturity. J. Information Systems Management 26, 138–152 (2009)
15. Lankhorst, M.: ArchiMate Language Primer. Final Project of Telematica Instituut (2004)

16. Liu, K., Sun, L., Tan, S.: ‘Modelling complex systems for project planning: a semiotics. J. International Journal of General Systems 35(3), 313–327 (2006)
17. Liu, K., Sun, L., Cook, S., Cui, G., Ketabchi, S., Neris, V., Qin, J.: Business aligned It Strategy (BAITS) - methodology and User Guide, Technical Report by collaboration between Capgemini and the University of Reading (2009)
18. Luftman, J.N.: Managing the Information Technology Resources. Pearson Prentice Hall, New Jersey (2004)
19. Pyburn, P.: Linking the MIS plan with corporate strategy: an exploratory study. MIS Quarterly 7(2), 1–14 (1983)
20. Qin, J., Liu, K., Han, J.: IT application valuation in business and IT alignment. In: The 2nd International Conference on Computer Engineering and Technology, Chengdu, pp. 186–190 (2010)
21. Qin, J., Liu, K., Han, J.: Service valuation in business and IT alignment: with a case study in banking in China. In: IEEE International Conference on Management of Innovation and Technology (ICMIT), Singapore, pp. 390–395 (2010)
22. Raghunathan, B., Raghunathan, T.S.: Planning System Success: Replication and Extension to the Information Systems Context. Working Paper, University of Toledo, Toledo, OH (1990)
23. Rood, M.A.: Enterprise Architecture: Definition, Content, and Utility. J. IEEE Trans. (1994)
24. Schekkerman, J.: How to Survive in the jungle of Enterprise Architecture Frameworks: Creating or choosing an Enterprise Architecture Framework, 2nd edn. Trafford Publishing, Canada (2004)
25. Schekkerman, J.: How are Organizations Progressing?. Web-form Based Survey, Report of the Third Measurement (2005), http://www.enterprise-architecture.info/EA_BP.htm
26. Sowa, J., Zachman, J.: Extending and formalizing the framework for information systems architecture. IBM Systems Journal 31(3), 590–616 (1992)
27. Spewak, S.H., Hill, S.C.: Enterprise Architecture Planning: Developing a Blueprint for Data, Applications, and Technology. John Wiley & Sons, New York City (1995)
28. Summers, K.: How to Make IT-Business Alignment a Reality. CIO Executive Council (2009), http://www.cio.com/article/500621/How_to_Make_IT_Business_Alignment_a_Reality
29. TOGAF 9.0, online guide (2009), http://www.togaf.org/togaf9/toc-pt1.html
30. Zachman International, http://www.zachmaninternational.com/index.php/home-article/89#maincol (accessed September 2009)
31. Zachman, J.: A framework for information systems architecture. IBM Systems Journal 26(3), 276–292 (1987)

The Specifications of the Weakly Hard Real-Time Systems: A Review

Habibah Ismail and Dayang N.A. Jawawi

Department of Software Engineering
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
81110 Skudai, Johor, Malaysia
habibahisma@gmail.com, dayang@utm.my

Abstract. A real-time system is one in which the temporal aspects of its behaviour are part of their specification. The problem with traditional real-time specification is no guarantees on when and how many deadlines may be missed can be given to the tasks because their specification is focused on met all the deadlines and cannot missed it, otherwise the tasks is totally failed. Thus, the weakly hard specification solve this problem with define a gradation on how the deadlines can be missed while still guaranteeing the tasks meets their deadlines. In this paper, a review has been made on the three specifications of real-time systems which is losses or missed of the deadlines can be permitted occasionally. Three criteria used in the evaluation are the process model, temporal specifications and predictability. These three criteria were chosen because the tasks in real-time systems are usually periodic in nature, have timing constraints like deadlines and the behaviour of the systems must be predictable. The three specifications we reviewed in this paper are the skip constraints known as skip factor *s*, (*m*,*k*)-firm deadlines and the weakly hard constraints. The objective of review is to find which specification is better in order to predict the behaviour of a task based on those three criteria. Based on our review, it is concluded that the weakly hard constraints outperforms the two conventional specifications of weakly hard real-time systems using that three criteria based on our evaluation by using a mobile robot case study due to its capability to specify in a clear of the distribution of deadlines met and missed.

Keywords: Weakly-hard real-time systems, the process model, temporal specifications, predictability, weakly-hard specifications.

1 Introduction

Real-time systems are computer systems that based on which the correctness of the results of a task is not only related to the computations performed, but also on the time factors at which the results are produced or outputs are generated. The traditional real-time systems are classified into two categories, one is hard real-time system and another is soft real-time system [1]. In applications of real-time, for hard real-time system, no deadline miss is tolerated or in other words, its deadline must meet successfully otherwise the task easy to face a failure. Meanwhile, for soft real-time

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 274–288, 2011.

system, deadline miss is tolerated but minimized and occasionally, however, the term occasional is not precise with meaningful. Nevertheless, it is still acceptable even though the task is delayed because missing a deadline usually happens in a non-predictable way. The new generation for a real-time system is weakly hard real-time system which is provides a mechanism that can tolerate some deadlines with specifies in a clear, predictable and bounded way where the deadlines can be missed [2].

The problem with traditional real-time specification is in hard real-time systems, it is very restrictive because all the tasks must meet their deadlines. Meanwhile, in soft real-time systems, it is too relaxed because no guarantees can be given to the deadlines either it is met or missed. As hard real-time and soft real-time systems are miss restriction and miss tolerate respectively, the weakly hard real-time system can integrate both of this requirements by distributed the way of missed and met deadlines accurately. In weakly hard real-time systems, missing some of the deadlines is allowed occasionally, however it is still necessary and crucial to finish the tasks within a given deadlines. Likewise, it is concerns the specification of real-time systems where a specified number of deadlines can be missed is more precisely defined.

There have been three available and well-known approaches for specifying real-time systems that can let some deadlines to be missed occasionally. Firstly, the skip factor, s which a task has a skip factor of s it will have one invocations skipped out of s [3], secondly, the notion of (m,k)-firm deadlines which is to specify tasks that are required to meet m deadlines in any k consecutive invocations [4] and lastly, the weakly hard constraints with the simplest one is the $\binom{n}{m}$ constraint that a task meets any n in m deadlines, at least n invocations that meet the deadlines [5]. The criterion for reviewing the specification is based on the process model, temporal specifications and predictability. These three criteria were chosen because in [3], [4] and [5], those criteria were mentioned and used in order to evaluate their method.

The reason to review these three specifications it is because they are known as most widely used in the sense of weakly hard real-time systems. Therefore, from the review, we want to identify most suitable specification based on the three criteria we used in this paper. The objective of this review is to find which specification is performing well due to be used in a scheduling analysis framework in order to predict the tasks performance.

This paper is organized as follows. In section 2 we review the related work. After that, the evaluation of the strategy and the criteria are in section 3. In section 4, three specifications of real-time systems that can tolerate losses and missed of deadlines occasionally are evaluated, which involve the skip factor, s, (m,k)-firm deadlines and weakly hard constraints. Section 5 and 6 will be discussing and concluding the result of the evaluation.

2 Related Work

There have been some grateful efforts were proposed by other researchers as previous works to the specification of the weakly hard real-time systems in different approaches.

The first work is done by [3], who introduced the skip over scheduling algorithms do skip some task invocations according to the notion of skip factor, s. If a task has a skip factor of s it will have one invocation skipped out of s consecutive invocations.

That means, the distance between two consecutive skips must be at least s periods (it can be specified as a (s – 1, s)-constraint). When s = infinity (∞), no skips are permitted. The advantage of using this specification is skipping a task instance has no effect on the release time of the next instance [3]. However, the disadvantage is a selected number of task invocations are discarded (or skipped) even though the tasks could meet their deadline or there may be available computation resources to finish on time.

Further, the notion of (*m,k*)-firm deadlines is a scheduling algorithm was initially introduced by [4]. It is characterized by two parameters: *m* and *k* which means at least *m* invocations meet their deadline in any *k* consecutive invocations of the task. The priority of a task is raised if it is close to not meeting *m* deadlines on the last *k* invocations. The advantage of using this specification is the requirements of real-time application can be more precisely expressed using the proposed deadline model [4]. However, the disadvantage is it does not provide any guarantees on the number of deadlines a task can be missed.

Recently, the $\binom{n}{m}$ constraint which is equivalent to the (*m*,*k*)-constraint was introduced by [5] with the other three constraints: $\overline{\binom{n}{m}}$, $\left\langle {n \atop m} \right\rangle$, $\overline{\left\langle {n \atop m} \right\rangle}$ and they called as weakly hard temporal constraint. The advantage of using this specification is it clearly specified bounds on the number of deadlines a task may miss and on the pattern how these deadlines can be missed. They summarize the temporal properties of specifications available of weakly hard real-time systems; point out the relations between various specifications. Furthermore, they point out that specification should be considered from two aspects [5]:

1) A task maybe is sensitive to the consecutiveness of deadline met while another is only sensitive to the number of deadline missed.
2) A task maybe is sensitive to the consecutiveness of deadline missed while another is only sensitive to the number of deadline missed.

Furthermore, the evaluation of (*m,k*)-firm deadlines is done by [7] with compared their work with (*k*,*k*)-firm constraint (or equivalently hard real-time) in order to show (*m,k*)-firm deadlines is more relaxed in terms of deadline concept. By using case study, they concluded that the (*m*,*k*)-firm deadlines outperforms the (*k*,*k*)-firm constraints in terms of relaxing the resource requirements.

These previous works have introduced the idea of systems which tolerate some missed deadlines and this makes it easy for us to generalise their ideas and therefore to be able to apply their work in a more general framework.

3 Evaluation of the Strategy and the Criteria for Weakly Hard Real-Time Systems

In this paper, we apply three criteria namely the process model, temporal specifications and predictability from [2] to evaluate the systems that may tolerate some degree of missed or losses deadlines, through the adoption of a real-time concepts. Figure 1 shows the process flow for our review about the criteria that we identified for weakly hard specifications.

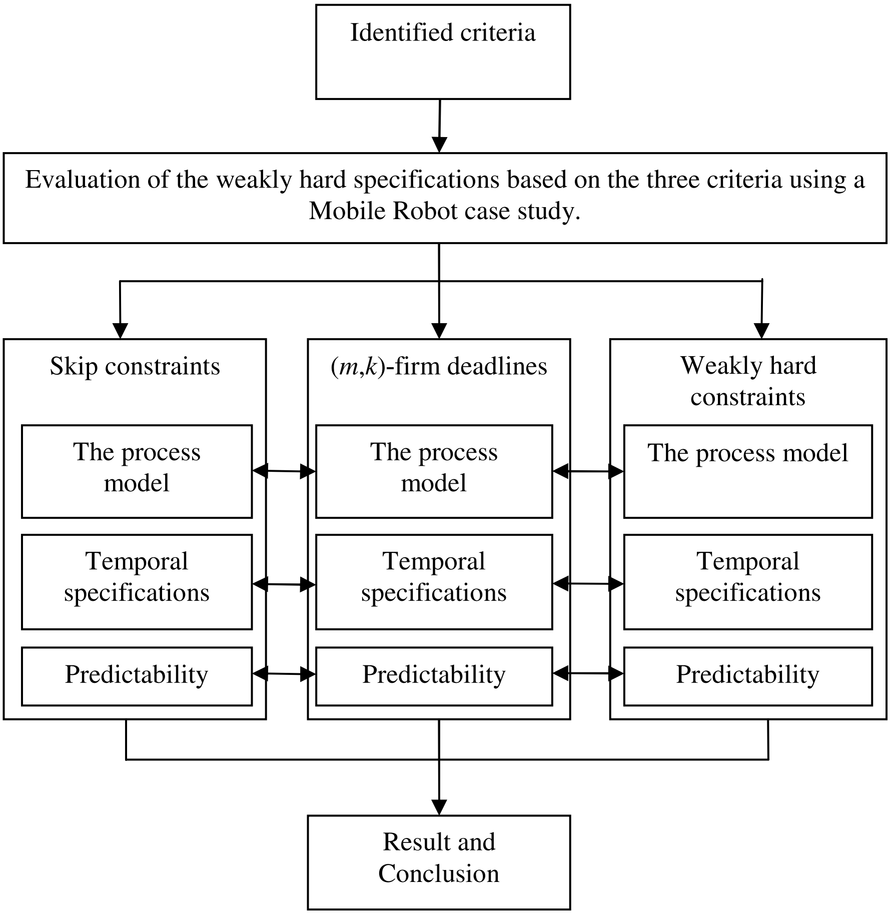


Fig. 1. The process flow of the review

In order to show the benefits of using these three criteria into the weakly hard specifications, a suitable task set was used. A Mobile robot case study was chosen because this case study consist two types of tasks, hard tasks and soft tasks [6]. Thus, it can be clarified as unnecessary to meet the entire task and message deadlines as long as the misses (or deadlines) are spaced distantly/evenly. The followings are descriptions and the importance on each of the criteria:

1) The process model is described as a periodic operation consists of a computation that is executed repeatedly, in a regular and cyclic pattern. These operations are usually called tasks and the tasks are independent. Formally, τ denotes a periodic task. The period of periodic tasks is denoted by T_i, the relative deadline of the task is denoted by D_i and N is the number of tasks. The process model is important in weakly hard real-time systems because in order to know a method is useful or not,

we should have a task and the task must have period and deadline. Additionally, the values of period and deadline of the tasks are used to predict the behaviour of the tasks. Furthermore, we can know what types of each task and it make an easy to analyse the tasks.

2) The temporal specifications of the system are set by means of temporal constraints. The most widely used temporal constraint is the deadline. The importance of temporal specifications in weakly hard real-time systems is as guidance to the tasks because how we could define missing some deadlines is tolerated if we not put timing constraints into a task. So, we need to specify temporal constraints in order to ensure the task not exceeds the deadlines. The requirements of real-time applications can be more precisely expressed using the temporal specifications.
3) The predictability has the property required to meet the timing requirements. Predictability requires that information about the system is known. The most important information needed is the worst case computation time of the tasks and it is denoted by C_i. The predictability is important due to guarantee the tasks meet all the deadlines and on the other hand, to ensure the tasks meet the temporal constraints. In weakly hard real-time systems, even though some deadlines can be missed, the tasks still can be guaranteed predictable because it is still be able to meet the timing requirements by specify in clear such a met and missed deadlines of the tasks.

These three criteria were chosen to reviewed in this paper because the tasks in real-time systems are usually periodic in nature, has deadlines and real-time systems commonly control systems. In control systems, the behaviour must be predictable. In fact, the notion of predictability and timing constraints like deadlines is very important to real-time systems.

Furthermore, each instance of a task has a deadline constraint by which it is expected to complete its computation [4]. In order to identify a most suitable specification, these three criteria are important where the process model that consists period and deadline allows an easy verification of the satisfaction of temporal specifications. Then, it is important to guarantee the systems meeting all the deadlines due to be make the systems be predictable properly.

4 Evaluation of the Result

After we discussed about the strategy and the criteria, each specification is evaluated in order to generate the results. The followings are the evaluation result of skip factor, s, (m,k)-firm deadlines and weakly hard constraints based on each criteria.

4.1 Skip Factor, s

The notion of skip factor, s is a task which has a skip factor of s it will have one invocation skipped out of s consecutive invocations [3]. That means that the distance between two consecutive skips must be at least s periods. That is, after missing a deadline at least $s - 1$ task instances must meet their deadlines [3]. When s equals to infinity ($s = \infty$), no skips are allowed.

4.1.1 The Process Model

The process model for skip factor *s* is represented by a task and each task is characterised by its period (T_i) and the deadline (D_i). We specifies the relative deadline of a task equals to its period ($D_i = T_i$). The worst case computation time C_i will be necessary for used in the schedulability tests. In additionally, each task is divided into instances where each instance occurs during a single period of the task. Every instance of a task can be *red* or *blue* [3]. A red task instance must complete before its deadline; a blue task instance can aborted at any time or in other words, task instance miss its deadline.

Therefore, the reader should realise the skip factor, *s* can fulfill the basic of the process model with additional information. Table 2 shows evaluation of skip factor, *s*, based on the first criteria, the process model.

Table 1. Evaluation based on the process model for skip factor, *s*

Task	**Period (T_i) (ms)**	**Deadline (D_i) (ms)**	**C_i (ms)**
MobileRobot	50	50	1
motorctrl_left	50	50	20
motorctrl_right	50	50	20
Subsumption	80	80	1
Avoid	100	100	17
Cruise	100	100	1
manrobotintf	500	500	16

From the table above, we specified motorctrl_left, motorctrl_right, Subsumption and Avoid task as a *red* task equivalent to a hard periodic task. In other words, no skips are allowed for these four tasks and the tasks must finish within its deadline. Meanwhile, for MobileRobot, Cruise and manrobotintf task, these three tasks can have *red* and *blue* task instance in each task depends on how this instance executed during a period of time. A blue instance has been rejected if its fail to complete before its deadline in order to satisfied a red instance to finish before its deadline. It is because the priority is given to red task instance.

4.1.2 Temporal Specifications

The possible skips of a task is characterised by its skip parameter $2 \leq s \leq \infty$, which represents the tolerance of this task to missing deadlines [3]. That means this parameters gives the distance between two consecutive skips must be at least *s* periods. The skip constraint is used to make spare time in order to ensure all tasks complete or finish within a given deadline.

Table 2 shows evaluation of skip factor, *s*, based on the second criteria, temporal specifications.

Table 2. Evaluation based on temporal specifications for skip factor, *s*

Task	Period (T_i) (ms)	Deadline (D_i) (ms)	Skip factor *s*
MobileRobot	50	50	2
motorctrl_left	50	50	∞
motorctrl_right	50	50	∞
Subsumption	80	80	∞
Avoid	100	100	∞
Cruise	100	100	2
manrobotintf	500	500	2

From the table above, consider as example, for motorctrl_left task until Avoid task, *s* is equal to infinity, that means these four tasks are not allowed to skip. For MobileRobot, Cruise and manrobotintf task, we specify that three tasks have the same skip parameters, $s = 2$. Hence, the tasks are permitted to skip one instance every 50, 100 and 500 periods but just for two times. The value of *s* can be any number depends on how many times we want each task to skip.

4.1.3 Predictability

Basically, two scheduling algorithms were used in the skip-over model. Firstly, the Red Tasks Only (RTO) algorithm is used to schedule red tasks instances according to Earliest Deadline First (EDF) algorithm. Secondly, the Blue When Possible (BWP) algorithm is used to schedule blue tasks instances without bother any red task from completing their execution within its deadlines.

In order to show the benefits of using the skip factor, *s* on the third criteria, predictability, a suitable task set was used. The following case study, based on the one described by [6] will be useful to illustrate that evaluation. Table 3 shows the task parameters of the task set.

Table 3. Evaluation based on predictability for skip factor, *s*

Task	Period (T_i) (ms)	Deadline (D_i) (ms)	C_i (ms)	Skip factor *s*
MobileRobot	50	50	1	2
motorctrl_left	50	50	20	∞
motorctrl_right	50	50	20	∞
Subsumption	80	80	1	∞
Avoid	100	100	17	∞
Cruise	100	100	1	2
manrobotintf	500	500	16	2

From the table above, we note that the seven tasks cannot be scheduled, because this task set have the processor utilization factor, U = 1/50 + 20/50 + 20/50 + 1/80 + 17/80 + 1/100 = 16/500 = **1.08** > **1** and consequently some instances will necessary miss their deadlines in order to guarantee all tasks complete or finish within a given deadline. We note that motorctrl_left, motorctrl_right, Subsumption and Avoid task are red instance and indeed are not allowed to skips. On the other hand, definitely no blue instance in these four tasks. Meanwhile, for MobileRobot, Cruise and manrobotintf task, it is possible to have red and blue task instance.

Figure 2 shows MobileRobot task is skip one instance at second execution in order to give spare time to other tasks to execute during a period of time due to guaranteed the task complete before their deadline. The reader should realise, motorctrl_left and motorctrl_right task succeed to finish before its deadlines.

On the other hand, the first occurrence for MobileRobot task is red, thus according to BWP algorithm, a higher priority is always assigns to red tasks instances, that means the first occurrence of MobileRobot task namely red instance must completed within its deadlines. However, the second occurrence for that task is blue instance, so missing its deadlines is allowed due to give the second occurrence for motorctrl_left task namely red instance to execute and completed within its deadlines.

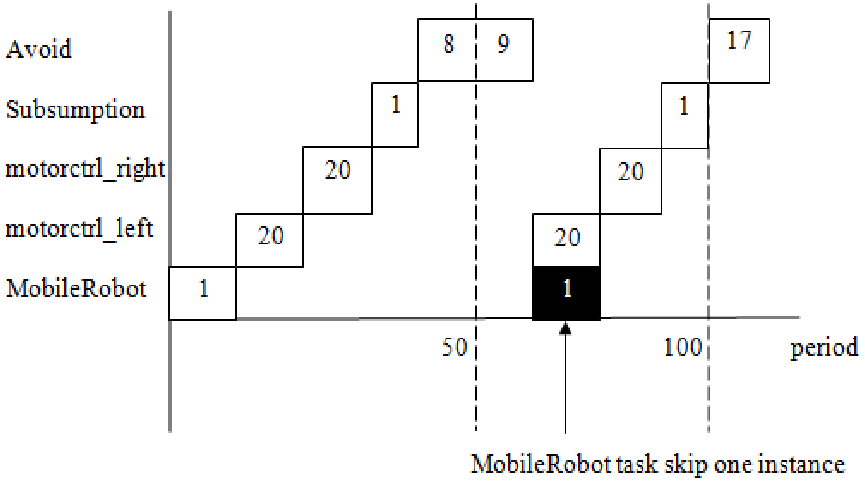


Fig. 2. Skipping a task

4.2 (*m,k*)-Firm Deadlines

The notion of (*m,k*)-firm deadlines is expand from the notion of skip factor whereas to specify tasks (or message) which are desired meets at least *m* deadlines in any window of *k* consecutive invocations [4]. Also, have the same *m* and *k* parameters for all tasks where it is declared as if the task is fully finish before or at the end of deadline, the task is successfully met its deadline. Otherwise, the task is missed it or dynamic failure occurs.

4.2.1 The Process Model

The process model for (*m*,*k*)-firm deadlines is represented by a task and each task is characterised by its period (T_i) and the deadline (D_i). We specifies the relative deadline of a task equals to its period ($D_i = T_i$).

Therefore, the reader should realise the (*m*,*k*)-firm deadlines can fulfill the basic of the process model.

4.2.2 Temporal Specifications

A task has two parameters *m* and *k* such as that a dynamic failure occurs if less than *m* out of *k* consecutive tasks meets their deadlines [4]. A task is said to have a (*m*,*k*)-firm guarantee if it is adequate to meet the deadlines of *m* out of *k* consecutive invocations of the task. In other words, task should meet *m* deadlines for every *k* consecutive invocations. The numbers *m* and *k* are integers such that $m \leq k$. The deadline model has two parameters *m* and *k* is used to better characterise the timing constraints. Table 4 shows evaluation of (*m*,*k*)-firm deadlines, based on the second criteria, temporal specifications.

Table 4. Evaluation based on temporal specifications for (*m*,*k*)-firm deadlines

Task	Period (T_i) (ms)	Deadline (D_i) (ms)	(*m*,*k*)
MobileRobot	50	50	(1,5)
motorctrl_left	50	50	(1,1)
motorctrl_right	50	50	(1,1)
Subsumption	80	80	(1,1)
Avoid	100	100	(1,1)
Cruise	100	100	(1,5)
manrobotintf	500	500	(2,5)

From the table above, it shows four from seven tasks has (1,1)-firm deadlines, namely motorctrl_left, motorctrl_right, Subsumption and Avoid which means those four tasks indeed must meet their deadlines. In a MobileRobot and Cruise task with (1,5)-firm deadlines, that means at least one invocations in any window of five task must meet their deadlines. Meanwhile, a manrobotintf task have (2,5)-firm deadlines corresponds to the constraint that no four consecutive invocations of the task should miss their deadlines.

4.2.3 Predictability

The prediction for the (*m*,*k*)-firm deadlines is using the Distance-Based Priority (DBP) scheme, by assign a priority to each task based on the recent of missed deadlines. Hence, a task that a closer to will not meeting its deadlines is assign higher priority in order to avoid tasks from dynamic failure and as consequently can raise its chances of meeting its deadlines.

In order to show the benefits of using the (*m*,*k*)-firm deadlines on the third criteria, predictability, a suitable task set was used. The following case study [6], will be useful to illustrate that evaluation. Table 5 shows the task parameters of the task set.

Table 5. Evaluation based on predictability for (*m*,*k*)-firm deadlines

Task	**Period (T_i) (ms)**	**Deadline (D_i) (ms)**	**(*m*,*k*)**	**Sequence**
MobileRobot	50	50	(1,5)	{00001}
motorctrl_left	50	50	(1,1)	{1}
motorctrl_right	50	50	(1,1)	{1}
Subsumption	80	80	(1,1)	{1}
Avoid	100	100	(1,1)	{1}
Cruise	100	100	(1,5)	{00010}
manrobotintf	500	500	(2,5)	{11000}

From the table above, under DBP scheme, manrobotintf task with (2,5)-firm deadlines is given a higher priority than MobileRobot and Cruise task with (1,5)-firm deadlines because it will closer to not meeting *m* deadlines on the last *k* invocations. Thus, to avoid manrobotintf task from failure, by using DBP scheme, we need to left shift the sequence and adding in the right side whether 1's which represent a deadline met or 0's which represent a deadline miss depends on the task missed or met its deadline. The following is the result after we left shifted the manrobotintf sequence and added with 1's in the right:

$$1\underline{1000} = 1000\underline{1} \tag{1}$$

By assign priorities to tasks based on the recent of missed deadlines, its chances of meeting its deadlines are raised and the probability of dynamic failure and probability of a task missing its deadline can be reduced.

4.3 Weakly Hard Constraints

The $\binom{n}{m}$ constraint is equivalent to the $\binom{m}{k}$-firm deadlines and known as weakly hard constraints. The consecutiveness of lost deadlines is very sensitive for some systems while others are only sensitive to the number of deadlines missed.

The merge of the two judgments, (a) consecutiveness vs. non-consecutiveness, and (b) missed vs. met deadlines concretely guides to four basic constraints ($n \geq 1$, $n \leq m$) [5]:

1) A task τ "meets any *n* in *m* deadlines", denoted $\binom{n}{m}$, if, in any window of *m* consecutive invocations of the task, there are at least *n* invocations in any order that meet the deadline.
2) A task τ "meets row *n* in *m* deadlines", denoted $\left\langle {n \atop m} \right\rangle$, if, in any window of *m* consecutive invocations of the task, there are at least *n* consecutive invocations that meet the deadline.

3) A task τ “misses any n in m deadlines”, denoted $\overline{\binom{n}{m}}$, if, in any window of m consecutive invocations of the task, no more of n deadlines are missed.
4) A task τ “misses row n in m deadlines”, denoted $\overline{\left\langle {n \atop m} \right\rangle}$, if, in any window of m consecutive invocations of the task, it is never the case that n consecutive invocations miss their deadline.

4.3.1 The Process Model

The process model for the weakly hard constraints is represented by a task. Each task is characterised by the period of the task, T_i, the deadline of the task, D_i and a weakly hard constraints, λ. We assume that $D_i = T_i$. The worst case computation time C_i will be necessary for used in the schedulability tests.

Therefore, the reader should realise the weakly hard constraints can fulfill the basic of the process model.

4.3.2 Temporal Specifications

The weakly hard constraints have four temporal constraints in order to model the tasks. One of them defined as a task has two parameters n and m such as that a task meets any n in m deadlines if in any window of m consecutive invocations of the task, there are at least n invocations that meet the deadline.

Table 6 shows evaluation of weakly hard constraints, based on the second criteria, temporal specifications.

Table 6. Evaluation based on temporal specifications for the weakly hard constraints

Task	**Period (T_i) (ms)**	**Deadline (D_i) (ms)**	**λ**
MobileRobot	50	50	<2,5>
motorctrl_left	50	50	(1,1)
motorctrl_right	50	50	(1,1)
Subsumption	80	80	(1,1)
Avoid	100	100	(1,1)
Cruise	100	100	$\overline{(1,5)}$
manrobotintf	500	500	$\overline{(1,5)}$

From the table above, the traditional assumptions that a task must meet its deadline can be represented as (1,1)-firm deadlines and it shows four from seven tasks has it. A MobileRobot task have <2,5>-firm deadlines corresponds to the constraint that the task has to meet at least two consecutive deadlines in any five consecutive invocations. Meanwhile, in a Cruise and manrobotintf task with $\overline{(1,5)}$-firm deadlines expresses, no more one deadlines are missed in any five consecutive invocations.

4.3.3 Predictability

In order to guarantee all tasks meets its deadlines and satisfied their weakly hard constraints, additional information is needed as a way to make the tasks predictable. In order to shows an exact number of deadlines met and can be missed, response times and hyperperiod analysis are used. The following formula is the equation for the response times, R_i [5]:

$$Ri = Ci + \sum_{\tau j \in hp(\tau i)} \left\lceil \frac{Ri}{Ti} \right\rceil Cj \tag{2}$$

The higher order period or the hyperperiod, h_i consist the number of invocations of a task in the hyperperiod, $A_i = \frac{H}{Ti}$ and the number of invocations of a task in the hyperperiod at level i, $a_i = \frac{hi}{Ti}$ [5].

The values of the hyperperiod are getting by using least common multiple tool. The equation for the calculation of the hyperperiod h_i is given by [5]:

$$h_i = lcm\{T_j \mid \in hep(\tau_i)\} \tag{3}$$

Table 7 shows evaluation of weakly hard constraints, based on the third criteria, predictability.

Table 7. Evaluation based on predictability for the weakly hard constraints

Task	T_i	D_i	C_i	R_i	h_i	a_i	A_i	**λ**
MobileRobot	50	50	1	1	50	1	40	(1,1)
motorctrl_left	50	50	20	21	50	1	40	(1,1)
motorctrl_right	50	50	20	41	50	1	40	(1,1)
Subsumption	80	80	1	42	400	5	25	(5,5)
Avoid	100	100	17	100	400	4	20	(5,5)
Cruise	100	100	1	160	400	4	20	(2,4)
manrobotintf	500	500	16	236	2000	4	4	(2,2)
					$\sum$	**20**	**189**	

From the table above, for Cruise task, it is means the task is invoked $\alpha_6 = 4$ times within the hyperperiod at level 6. In the worst case only two invocations out of four will miss the deadline. Thus, the weakly hard constraint for Cruise task is defined as $\binom{2}{4}$ constraint.

From the result, using weakly hard constraints, allow us to model tasks like this one with specify bounds on the number of deadlines a task may miss and on the pattern of how these deadlines can be missed with find the hyperperiod for each task. Therefore, we can know how many times each task is invoked within the hyperperiod and from the invocations, we can get the number of deadlines missed and met for each task. Most importantly, the tasks still guaranteed to finish within a given deadlines although there have one task miss its deadlines in the worst case because the task satisfies its temporal constraints and indeed missed some of the deadlines is acceptable in weakly hard real-time systems.

5 Discussion and Summary of Evaluation

The evaluation of specifications of real-time systems based on their criteria is summarized in Table 8. In this paper, comparisons have been made on the three specifications, with respect to those three criteria.

Table 8. Summary of evaluation

Specification / Criteria	Skip factor, s	(m,k)-firm deadline	Weakly hard constraints
The Process Model	Task, period, deadline, red or blue task instance	Task, period and deadline	Task, period and deadline
Temporal Specifications	Skip constraint	Deadline model	Four temporal constraints
Predictability	The value of s	Sequence denoted by ones and zeroes	R_i, The hyperperiod h_i, A_i and a_i

The following discussion and conclusion are based on each criterion. For the process model, the skip factor, s have task, period, deadline and each task can be either red or blue. For the notion (m,k)-firm deadlines and the weakly hard constraints, both methods have the same basic process model due to its had a task and each task is characterised by period and deadline. Thus, it makes skip factor, s is different and good from another two methods in this criteria because by specifies a task to be red or blue instance, can makes it easy to determine which task is allowed to skip and which is not because each task was identified at first.

For temporal specifications, the skip factor, s have been used skip constraints which means, some skip is allowed to certain task in order to ensure all tasks complete or finish within a given deadlines. For the notion (m,k)-firm deadlines, the deadline model is used in order to express the number of deadline met and missed by define a task with two parameters, m and k. Meanwhile, for the weakly hard constraints, this constraint have two parameters n and m and it is improvement and equivalent from the (m,k)-firm deadlines and that constraint can define how many deadlines can be missed in a single task with precisely and clearly. It is because those methods have four different temporal constraints in order to specify the tasks. Also, those temporal constraints can model the tasks such as the met or missed deadline is happens in consecutive or non-consecutive. As conclusion, the skip constraints is useful only in order to guarantee all tasks meets their deadlines, by gives skips to some task. Nevertheless, this method does not provides any guarantees for the tasks to meet its temporal constraints but only to do skips based on the value of skip factor, s. While, another two methods are concerns on to guarantee all the tasks meets its deadlines and also to satisfy its temporal constraints. The similarity between both specifications is because the expression for one of the weakly hard temporal constraint namely $\binom{n}{m}$-constraints is an approximately same with (*m*,*k*)-firm deadlines. However, the weakly hard constraints is better than the (*m*,*k*)-firm deadlines due to the fact it have four temporal constraints and definitely four constraints is much better and detailed compared with one constraint in order to specify the tasks.

For predictability, the skip factor, *s* can be used to guarantee the tasks meet its deadlines by allowed skips for a task. Nevertheless, it does not make any sense in order to know exactly the amount of deadlines met and missed. It is because skip constraint is used only to allow some task to skip in order to ensure the tasks finish before the deadlines. The value of *s* is helpful as guidance (or limitation) to the task because it consist the number of times that the task can be skipped within a given period. For the notion (*m*,*k*)-firm deadlines, the prediction is based on assign priorities to tasks. In order to satisfy their temporal constraints, a task that is closer to not meeting its deadline is given a higher priority due to avoid dynamic failure. Meanwhile, for the weakly hard constraints, it can predict much better from other two because this method specified in clear on the number of deadlines met and can be missed with define whether the task met and missed its deadlines consecutive or non-consecutive. With known in a clear where these deadlines are met and missed can be ruled for the tasks to meet its temporal constraints.

As conclusion, we can say that the weakly hard constraints fulfill all three criteria with high expectations because based on our discussion above, the basic process model is used as well as to give enough information to that method. Instead of temporal specifications, the weakly hard constraints can model the tasks with specifies in a clear and precise about the consecutiveness of deadline met and missed. In order to guarantee the tasks are predictable, the satisfaction of temporal constraints should taken importantly and indeed all the tasks must meet their deadlines by not finish over a given deadlines.

6 Conclusion

In conclusion, we can conclude that the weakly hard constraint is useful to be a beneficial specification because it allows us to predict what the behaviour of a task if a deadline is missed. This is due to the fact that from the comparison, it shows the weakly hard constraints outperforms the skip factor, *s* and the (*m*,*k*)-firm deadlines because the method can fulfill those three criteria namely the process model, temporal specifications and predictability very well after we illustrated with a mobile robot case study. A weakly hard constraint is valuable on how to model tasks with met and missed deadlines whenever happen to the tasks. Likewise, this method concerned about a met and a miss the deadlines of a task whether as consecutive or non-consecutive and also on how many deadlines met and missed for each task. Thus, the weakly hard constraints appear as most suitable specifications based on the three criteria we mentioned before.

For future work, our aim is to integrate the weakly hard requirements with MARTE modeling language in order to support our scheduling analysis framework. After working with the specifications, our continued work is on the analysis of weakly hard real-time systems. As additional information for the reader, the weakly hard constraint have a μ-patterns, in which this pattern is based on to characterise a met and missed deadlines of task by a binary sequence and we want to know is it this pattern can be used into (*m*,*k*)-firm deadlines because both of specifications has some similarities.

Acknowledgments. The authors thank profoundly to the Fundamental Research Grant Student (FRGS), the Software Engineering Research Group (SERG), the Universiti Teknologi Malaysia (UTM) and not forgotten, EReTSEL lab members for their support and help.

References

1. Shin, K.G., Ramanathan, P.: Real-time computing: A new discipline of computer science and engineering. Proceedings of the IEEE 82(1) (1994)
2. Bernat, G., Burns, A.: Weakly hard real-time systems. IEEE Transactions on Computers 50(4), 308–321 (2001)
3. Koren, G., Shasha, D.: Skip-over: algorithms and complexity for overloaded systems that allow skips. In: Proceedings of the 16th IEEE Real-Time Systems Symposium, Pisa, Italy, pp. 110–117 (December 1995)
4. Hamdaoui, M., Ramanathan, P.: A dynamic priority assignment technique for streams with (m,k)-firm deadlines. IEEE Transactions on Computers 44(12), 1443–1451 (1995)
5. Bernat, G.: Specification and analysis of weakly hard real-time systems. PhD Thesis, Department de les Ciències Matemàtiques i Informàtica. Universitat de les Illes Balears, Spain (January 1998)
6. Jawawi, D.N.A., Deris, S., Mamat, R.: Enhancement of PECOS embedded real-time component model for autonomous mobile robot application. In: IEEE International Conference on Computer Systems and Applications, pp. 882–889 (2006)
7. Li, J., Song, Y.-Q., Francoise, S.-L.: Providing real-time applications with graceful degradation of Qos and fault tolerance according to (m,k)-firm model. IEEE Transactions on Industrial Informatics 2(2), 112–119 (2006)
8. Bernat, G., Cayssials, R.: Guaranteed on-line weakly hard real-time systems. In: 22nd IEEE Real-Time Systems Symposium (RTSS 2001), London, England, pp. 25–35 (December 2001)
9. Broster, I., Bernat, G., Burns, A.: Weakly hard real-time constraints on controller area network. In: 14th Euromicro Conference on Real-Time Systems (ECRTS 2002), Vienna, Austria, p. 134 (June 2002)
10. Silly-Chetto, M., Marchand, A.: Dynamic scheduling of skippable periodic tasks: issues and proposals. In: 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based (ECBS 2007), Tucson, Arizona, USA, pp. 171–177 (March 2007)
11. Wang, Z., Song, Y.-Q., Poggi, E.-M., Sun, Y.: Survey of weakly hard real-time schedule theory and its application. In: Proceedings International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), pp. 429–437 (2002)
12. Wang, S.-H., Tsai, G.: Specification and timing analysis of real-time systems. Real-Time Systems 28(1), 69–90 (2004)

Extracting Named Entities from Prophetic Narration Texts (Hadith)

Fouzi Harrag[1], Eyas El-Qawasmeh[2], and Abdul Malik Salman Al-Salman[3]

[1] Compt Sci Department, College of Science,
Farhat ABBAS University,
Setif, 19000, Algeria
hfouzi2001@yahoo.fr
[2, 3] College of computer & Information Science,
King Saud University,
Riyadh, 11543, Saudi Arabia
{eelqawasmeh,salman}@ksu.edu.sa

Abstract. In this paper, we report our work on a Finite State Transducer-based entity extractor, which applies named-entity extraction techniques to identify useful entities from prophetic narrations texts. A Finite State Transducer has been implemented in order to capture different types of named entities. For development and testing purposes, we collected a set of prophetic narrations texts from "*Sahîh Al-Bukhari*" corpus. Preliminary evaluation results demonstrated that our approach is feasible. Our system achieved encouraging precision and recall rates, the overall precision and recall are 71% and 39% respectively. Our future work includes conducting larger-scale evaluation studies and enhancing the system to capture named entities from chains of transmitters (*Salasil Al-Assanid*) and biographical texts of narrators (*Tarajims*).

Keywords: Text Mining, Information Extraction, Named entity Extraction, Prophetic Narrations Texts, Finite State Transducer.

1 Introduction

The information age has made it easy for us to store large amounts of texts. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, while the amount of information available to us is constantly increasing, our ability to absorb and process this information remains constant.

The areas of information retrieval (IR) and information extraction (IE) are the subject of active research for several years in the community of Artificial Intelligence and Text Mining. With the appearance of large textual corpora in the recent years, we felt the need to integrate modules for information extraction in the existing information retrieval systems. The processing of large textual corpora leads needs that are situated at the border of information extraction and information retrieval areas. [10].

Information Extraction is perhaps the most prominent technique currently used in text mining pre-processing operations. Without IE (Information Extraction)

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 289–297, 2011.

techniques, text mining systems would have much more limited knowledge discovery capabilities [5].

As a first step in tagging documents for text mining systems, each document is processed to find (i.e., extract) entities and relationships that are likely to be meaningful and content-bearing. With respect to relationships, what are referred to here are facts or events involving certain entities. The extracted information provides more concise and precise data for the mining process than the more naive word-based approaches such as those used for text categorization, and tends to represent concepts and relationships that are more meaningful and relate directly to the examined document's domain.

This remainder of the paper is organized as follows: Section 2 presents related works on named entity extraction. In Section 3, we present definition of the finite-state transducers and their application in the domain of natural language processing. In Section 4, we present our approach for the named entity extraction in the corpus of “Sahih of Bukhari”. Experiences and Evaluation results are reported in Section 5 and finally Section 6 concludes this paper.

2 Related Works on Named Entity Extraction

Named Entity Extraction (NEE) task consists of detecting lexical units in a word sequence, referring to concrete entities and of determining which kind of entity the unit is referring to (persons, locations, organizations, etc.). This information is used in many NLP applications such as Question Answering, Information Retrieval, Summarization, Machine Translation, Topic Detection and Tracking, etc., and the more accurate the extraction of Named Entities (NE) is, the better the performance of the system will be [14]. The Message Understanding Conference (MUC) series has been the major forum for researchers in this area, where they meet and compare the performance of their entity extraction approaches.

Several successful systems for large-scale, accurate named entity recognition have been built. The majority of the systems operates on English text and follows a rule based and/or probabilistic approach, with hybrid processing being the most popular [6].

NEs in Arabic are particularly challenging as Arabic is a morphologically-rich and case-insensitive language. NE Recognition in many other languages relies heavily on capital letters as an important feature of proper names. In Arabic there is no such distinction. Most of the literature on Arabic NEs concentrates on NE recognition; NE Extraction is viewed largely as a subset of the task of NE Recognition. Benajiba et al., [4] presented ANERsys system built for Arabic texts, based on n-grams and maximum entropy. Their results showed that the used approach allows tackling the problem of NER for the Arabic language. Traboulsi [17] discussed the use of corpus linguistics methods, in conjunction with the local grammar formalism, to identify patterns of person names in Arabic news texts. The results show that the consideration of all the expectations related to the universal grammars could save a lot of time and effort needed. Elsebai et al., [9] described the implementation of a person name named entity recognition system for the Arabic Language. They adopted a rule based approach and used a set of keywords to reach the phrases that probably include person names. Their system achieved an F-measure of 89%. Attia et al., [3] adapted a Multilingual Named Entity Lexicon approach to Arabic, using Arabic WordNet (AWN) and Arabic Wikipedia (AWK). They achieved precision scores from 95.83% (with 66.13% recall) to 99.31% (with 61.45% recall) according to different values of a threshold. Shaalan et al., [16] developed the system, Name Entity Recognition for Arabic (NERA), using a rule-based approach. NERA is evaluated using semi-tagged corpora and the performance results achieved were satisfactory in terms of precision, recall, and f-measure.

3 Finite-State Transducers

Finite-State Transducers have been developed for a few years to parse natural language. The advantages of transducers are their robustness, precision and speed.

A transducer is a synchronous sequential machine with output; it is defined as follows: a synchronous sequential machine M is a 5-tuple, with $M = (I, O, S, fs, fo)$, where:

- I, O are finite non-empty input and output sets,
- S is a non-empty set of states,
- The function f_s: $I \times S \rightarrow S$ is a state transition mapping function which describes the transitions from state to state on given inputs,
- The function f_o: $S \rightarrow S$ is an output function.

While a finite state automaton is a system that either accepts or rejects a specific sequence, a transducer on the other hand transforms or "transduces" the sequence into a different output representation.

For some time, FST technology has been successfully used in morphological analysis and surface oriented syntactic natural language processing (NLP). Various models and formalisms have been proposed and implemented. Indeed, lots of software packages for FSTs exist, but the field is fragmented and partitioned which slows down the progress and the reuse of existing results.

Finite state techniques have been utilized on various tasks of computational linguistics before they have not been used in the domain of entity recognition. The NYU system for MUC-6 [8][15] uses sets of regular expressions which are efficiently applied with finite state techniques. The F-measure is 80%. The NERC system developed in DFKI [13] for German text processing is based on FST's and performance ranges between 66% and 87% for different NE types. Abney [1] presents a syntactic parser for texts in English or German language (Cass System). The system is based on a sequence of transducer to find the reliable patterns in the first stage and the uncertain patterns in the second stage. In the same way [12] presented a finite-state

Transducers to parse Swedish, which is very close to Abney's one. The IFSP System (Incremental Finite State Parser [2], created at Xerox Research Center) is another system of Finite transducers which has been used for a syntax analysis of Spanish language [7]. Fastus [11] is a very famous system for information extraction from texts in English or Japanese. This system parses texts into larger phrases to find locations, dates and proper names. Fastus is based on the use of Finite state transducer.

4 Our Approach for Named Entity Extraction

Our work in this paper, focus on the extraction of the surface information, i.e. information that not requires complex linguistic processing to be categorized. The goal is to detect and extract passages or sequences of words containing relevant information from the prophetic narrations texts. We propose Finite state transducers-based system to solve successively the problem of texts comprehension.

4.1 Corpus of "Sahîh of Bukhari"

Prophetic narrations texts or Hadith are considered as a very important source of Islamic legislation. Despite the importance of these collections, no reference corpus has been developed for research purposes. In our work, we opted for the use of a selection of prophetic narration texts from the "Sahîh of Bukhari" collection. "*Sahîh of Bukhari,* (صحيح البخاري)" is compiled by Bukhari scholar; it represents his most famous work. This book covers almost all aspects of life in providing proper guidance of Islam such as the method of performing prayers and other actions of worship directly from the Islamic prophet Muhammad. Bukhari organized his book as taxonomy of prophetic narration concepts which consists of three hierarchical levels. The first level is the one of the books; the number of books in "Sahîh of Bukhari" is 91. The second level is the one of the chapters, the number of chapters is 3882 and the third level is the one of the prophetic narration texts. Bukhari habitually took a portion of the Hadith for the heading of the chapter. Also he repeated the hadiths time after time according to their legal deductions. The number of hadiths in his book is 9,082 but without repetition it goes down to 2,602 [18].

4.2 Entity Extraction Task

The corpus on which we conducted our tests is the collection "*Sahîh of Bukhari*". The task consists of extracting the relevant conceptual information from the global text file of this corpus. More specifically, the system must detect the relevant text areas and assign a conceptual label among a finite set of labels: (*Num-Kitab, Title-Kitab, Num-Bab, Title-Bab, Num-Hadith, Saned, Matn, Taalik, and Atraf*). For each detected concept, the goal is to fill a pattern that contains a description of the event. The adopted approach allows us; first, to eliminate as soon as possible the not relevant information, and second, to successively refine the selection of variables (or terms) needed to detect the relevant information.

4.3 Finite Transducer Model

The extraction of the structural information from the prophetic text requires a dynamic treatment of the words in a passage and taking into account its context. For this purpose, we used a model of finite state transducer in the form of an automaton as described in Figure 1. This automaton convert a sequence of vectors (here: word code's) to a sequence of symbols (here: concepts that we want to extract), which correspond for example to the most likely set of labels for this sequence. The automatic model of the entity extraction is defined by a set of states and transitions between these states, and a coding text is associated to each state. In our model, the different states of the automaton encode the concepts, the transitions structure codes the automaton of concepts i.e. transitions between acceptable concepts.

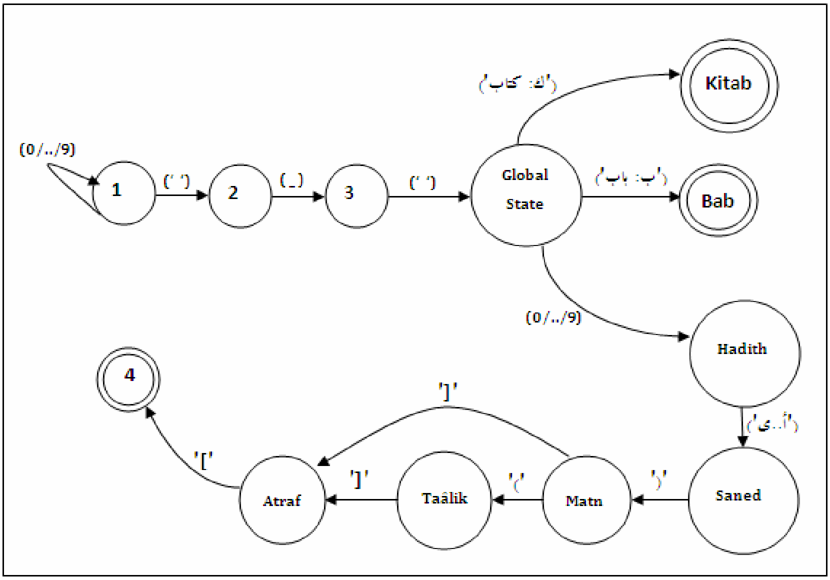


Fig. 1. Automaton of concepts for the sequences production in our named entity extraction

The entity extraction is the first step in the process of text mining in the corpus of prophetic narration texts. This operation will allow us to convert the unstructured text file of the '*Sahîh of Bukhari*' into a semi-structured text file. The extracted information will be used to annotate the prophetic narration texts of our corpus. The annotated corpus is considered as an initial source for the process of text mining. The extraction of the surface information will allow distinguishing the different parts (fields) of a Hadith (*Kitab, Bab, Saned, Matn* ... etc.). The main steps of the named entity extraction algorithm are detailed in algorithm 1.

Algorithm 1. Named entity extraction in the corpus of "Sahîh of Bukhari."

```
// 1. Initial state:
If character is digit {Car = 0…9} then
   Recognition of the entity:
     Book Number: {Num-Kitab}|
     Chapter number: {Num-Bab}|
     Hadith number: {Num-Had};
// 2. Extraction of the entity Book "Kitab-Title":
If character = "ك" then
   Recognition of the entity:
     Book Title: {Title-Kitab};
// 3. Extraction of the entity Chapter "Bab":
If character = "ب" then
   Recognition of the entity:
     Chapter Title: {Title-Bab};
// 4. Hadith Global State:
If character • "ك" and character • "ب" then
   // 5. Extraction of the entity "Saned":
   If character = '(' then
      Recognition of the entity: {Saned};
   // 6. Extraction of the entity "Matn":
   If character = ')' then
      Recognition of the entity: {Matn};
   // 7. Extraction of the entity "Taalik":
   If character = '[' then
      Recognition of the entity: {Taalik};
   // 8. Extraction of the entity "Atraf":
   If character = ']' then
      Recognition of the entity: {Atraf};
```

5 Experiments and Results

We used precision, recall and F_1 rates to measure the performance of our system. They were calculated as follows:

$$\text{Precision} = \text{Number of correct entities extracted by system} / \text{Number of all entities extracted by system} \quad (1)$$

$$\text{Recall} = \text{Number of correct entities extracted by system} / \text{Number of total entities extracted by human} \quad (2)$$

$$\text{F1-Measure} = (2 * \text{Recall} * \text{Precision}) \;/\; (\text{Recall} + \text{Precision}) \quad (3)$$

The results obtained by our system are summarized in table 1.

Table 1. Experimental results

NE Type	Precision	Recall	F1
Num-Kitab	1.00	0.50	0.67
Title-Kitab	1.00	0.50	0.67
Num-Bab	0.81	0.57	0.67
Title-Bab	0.94	0.65	0.77
Num-Had	0.76	0.42	0.54
Saned	0.44	0.26	0.33
Matn	0.61	0.35	0.45
Taalik	0.00	0.00	0.00
Atraf	0.82	0.48	0.61
Total	**0.71**	**0.39**	**0.52**

The performances of the system for each NE type are shown in Figure 2.

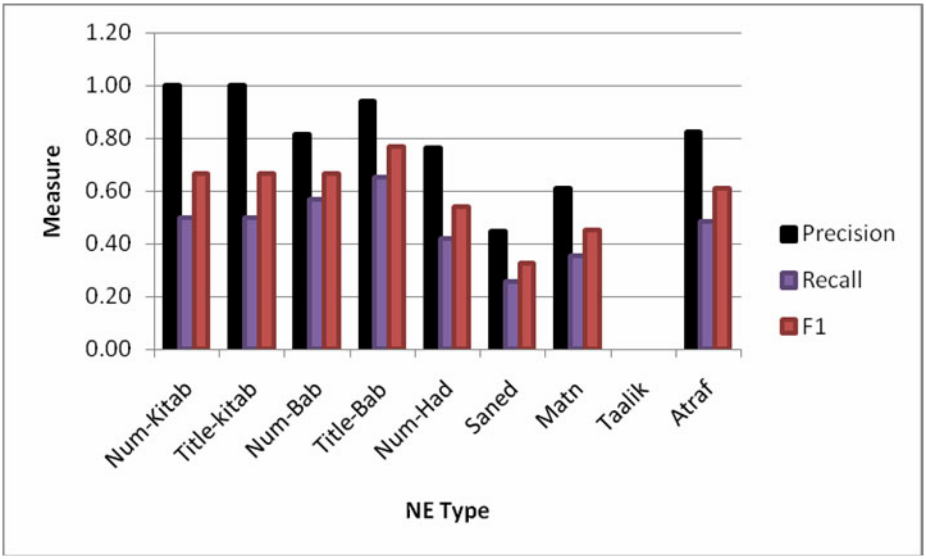


Fig. 2. Performance of "Hadith" named entity extraction

In general, the results indicate that our entity extractor performs well in identifying number entities like Num-Kitab, Num-Bab and Num-Hadith from our data set. We were satisfied considering that the prophetic narration texts were much noisier than the news articles used in MUC evaluation. However, the entity extractor's performance for "*Taalik*" entity was not as good as we expected. After examining the test data, we found that the system was unable to distinguish the "*Taalik*" information from the "*Matn*" information due to the textual nature of the two entities. We believed that the performance could be improved if this error were fixed.

6 Conclusion

In this paper, we report our work on applying named-entity extraction techniques to identifying useful entities from prophetic narration texts. We implemented finite state techniques favoring efficient text processing allowing fast customization to the needs and particularities of our specific application. Preliminary evaluation results have demonstrated the feasibility and the potential values of our approach. We observed the best results for the named entities Num-Kitab, Num-Bab and Num-Had with an F-measure of 0.67, 0.67 and 0.54 respectively, and explained this result by a relatively high ability of our system to detect digit entities. Our future work includes conducting larger-scale evaluation studies and enhancing the system to capture named entities from chains of transmitters (Salasil Al-Assanid) and biographical texts of narrators (Tarajims). We plan to improve our Finite state transducers-system by fixing some noted errors, and we are also interested by the use and the comparison of some others machine learning techniques for the task of named entity extraction.

References

1. Abney, S.: Partial parsing via finite-state cascades. In: Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic, pp. 8–15 (1996)
2. Ait-Mokhtar, S., Chanod, J.: Incremental finite state parsing. In: ANLP 1997 (1997)
3. Attia, M., Toral, A., Tounsi, L., Monachini, M., Genabith, J.V.: An automatically built Named Entity lexicon for Arabic, pp. 3614–3621 (2010)
4. Benajiba, Y., Rosso, P., BenedíRuiz, J.M.: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 143–153. Springer, Heidelberg (2007)
5. Ben-Dov, M., Feldman, R.: Text Mining And Information Extraction. In: Data Mining And Mvowledge Discovery Handbook, ch. 38, pp. 801–831. Mdx University, London (2006)
6. Demiros, I., Boutsis, S., Giouli, V., Liakata, M., Papageorgiou, H., Piperidis, S.: Named Entity Recognition in Greek Texts (2000)
7. Gala-Pavia, N.: Using the Incremental Finite-State Architecture to create a Spanish Shallow Parser. In: Proceedings of XV Congres of SEPLN, Lleida, Spain (1999)
8. Grishman, R.: The NYU system for MUC-6 or where's the syntax. In: Proceedings of Sixth Message Understanding Conference (1995)
9. Elsebai, A., Meziane, F., Belkredim, F.Z.: A Rule Based Persons Names Arabic Extraction System. Communications of the IBIMA 11, 53–59 (2009)
10. Harrag, F.: A text mining approach based on topic classification and segmentation Application to the corpus of Prophetic Traditions (Hadith), PhD thesis, Computer Science Dept., Faculty of Sciences, Farhat Abbas University, Setif, Algeria (2011)
11. Hobbs, J.R., Appelt, D.E., Bear, J., Israel, D., Kameyama, M., Stickel, M., Tyson, M.: FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In: Finite-State Devices for Natural Language Processing. MIT Press, Cambridge (1996)

12. Kokkinakis, D., Johansson-Kokkinakis, S.: A Cascaded Finite-State Parser for Syntactic Analysis of Swedish. In: Proceedings of the 9th EACL, Bergen, Norway (1999)
13. Neumann, G., Backofen, R., Baur, J., Becker, M., Braun, C.: An information extraction core system for real world German text processing. In: ACL (1997)
14. Padro, M., Padro, L.: A Named Entity Recognition System based on a Finite Automata Acquisition Algorithm. TALP Research Center, Universitat Politecnica de Catalunya (2005)
15. Pazienza, M.T. (ed.): SCIE 1997. LNCS (LNAI), vol. 1299. Springer, Heidelberg (1997)
16. Shaalan, K., Raza, H.: Arabic Named Entity Recognition from Diverse Text Types. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 440–451. Springer, Heidelberg (2008)
17. Traboulsi, H.: Arabic Named Entity Extraction: A Local Grammar-Based Approach. In: Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 139–143 (2009)
18. Wikipedia, Sahih Bukhari, http://fr.wikipedia.org/wiki/Sahih_al-Bukhari (last Visited May 30, 2011)

Security as a Service for User Customized Data Protection

Kenichi Takahashi[1], Takanori Matsuzaki[2], Tsunenori Mine[3],
Takao Kawamura[1], and Kazunori Sugahara[1]

[1] Department of Information and Electronics,
Graduate School of Engineering, Tottori University,
4-101 Koyama-Minami, Tottori, 680-8552, Japan
{takahashi,kawamura,sugahara}@isit.or.jp
[2] Kinki University School of Humanity-Oriented Science and Engineering,
11-6 Kashiwanomori, Iizuka-shi, Fukuoka 820-8555, Japan
takanori@fuk.kindai.ac.jp
[3] Faculty of Information Science and Electrical Engineering, Kyushu University,
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
mine@al.is.kyushu-u.ac.jp

Abstract. Some of Internet services require users to provide their sensitive information such as credit card number, and an ID-password pair. In these services, the manner in which the provided information is used is solely determined by the service providers. As a result, even when the manner in which information is used by a service provider appears vulnerable, users have no choice but to allow such usage. In this paper, we propose a framework that enables users to select the manner in which their sensitive information is protected. In our framework, a policy, which defines the type of information protection, is offered as a *Security as a Service*. According to the policy, users can incorporate the type of information protection into a program. By allowing a service provider to use their sensitive information through this program, users can protect their sensitive information according to the manner chosen by them.

1 Introduction

Nowadays, we use various Internet services in our business and daily life. Some of these services require users to furnish sensitive information such as their name, address, credit card number, and an ID-password pair when they use these services. For example, an online shopping site may request a user's address and credit card number for the shipping and the payment of an item. Further, cases of information leakage, such as those involving Yahoo! BB and CardSystems Solutions, have been reported and information leakage has now become a serious problem. For example, a leaked credit card number may lead to the loss of money and a leaked name may lead to privacy issues. Therefore, it is imperative that we exercise caution when providing sensitive information to Internet services. Therefore, a framework is required that will enable users to determine the manner in which sensitive information is protected.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 298–309, 2011.

We use a number of techniques such as encryptions, digital signatures and public key infrastructure (PKI) to secure sensitive information. These techniques are effective in protecting sensitive information from malicious third parties. However, they are not effective against the misuse of the information provided by users to legitimate service providers. Once service providers obtains sensitive information from users, there is a possibility that this information can be misused (e.g., leak, abuse, or accident), thus, resulting in information leakage. These are caused by a theft of the notebook PC, carelessness of the employees, illegal information carrying with a USB storage, and so on [1]. These problems may be solved by sufficiently educating employees, denying PCs access to external storage devices, and so on. However, even if the service providers take severe countermeasures, there is no way for the users to verify whether these countermeasures have been applied.

Security solutions using cloud computing technologies are attracting considerable attention. Because a number of services are expected to be implemented using cloud computing, all these services are collectively called *everything as a service (XaaS)*. In XaaS, if X stands for *security*, then it implies that security services are offered by cloud service providers. A malware detection service and a Web/URL filtering service offered by Mcafee[2] and Panda Security[3], respectively, are example of *Security as a Service (SaaS)*[1]. These services carry out malware detection and URL filtering on the cloud instead of using softwares installed on the client machine. Thus, users do not need to update the malware signatures and the URL blacklist. In addition, Yahoo! Japan and Rakuten Ichiba offer anonymous trading services [4, 5][2] that a buyer (user) can use to buy an item from a seller (service provider) without providing a name, an address, and a phone number to the seller, and vice versa. The buyers/sellers, however, have to provide sensitive information about themselves to an anonymous trading service provider instead of to the sellers/buyers. Therefore, it is still required that we confirm whether the anonymous trading service providers are enforcing strict measures to protect the information being provided to them, but there is no effective way to do this.

In this paper, we propose a framework that enables users to select the manner in which their sensitive information is protected. In this framework, a policy, which defines the type of information protection, is offered as a SaaS. The policies are availabe in the form of text-based information and open to the public. Because anyone can view and verify the policies, those policies that are used frequently can be considered to be reliable. A policy contains rules for replacing a part of a program in a manner so as to enable the protection of information. The user selects a policy that defines the manner in which his/her information is to be protected and the type of information protection defined by this policy is incorporated into a program. We call such a program a *customized program*. By allowing a service provider to the sensitive information provided by a user through the customized program, the user can protect his/her sensitive information in a manner selected by him/her. Consequently, both the service provider and the user can determine the manner in which information is used so as to protect the user's information. Moreover, this reduces the responsibility of

[1] SaaS usually means *Software* as a Service, but this paper uses SaaS for *Security* as a Service.

[2] They are not called as SaaS, but they are actually a kind of SaaS.

service providers as far as information protection is concerned, because user would also be involved in determining the manner in which his/her information is protected.

The remainder of this paper is structured as follows. Section 2 describes the related studies. Section 3 presents our framework. Then, we describe an application scenario in Section 4. Finally, Section 5 concludes the paper.

2 Related Studies

Cryptographic algorithms such as symmetric and public-key algorithms as well as techniques based on them such as digital signatures and public key infrastructure (PKI) have been proposed [6]. These algorithms and techniques aim at preventing message interception or identification of communication partners. Thus, they ensure message confidentiality, integrity, and availability of the message against malicious exploitation by third parties. Such techniques, however, do not prevent communication partners from abusing sensitive information released to them.

We often find a link on certain websites that points to the privacy policy adopted by that websites; this is termed as *Privacy Policy* on Yahoo!, *Privacy* on IBM, and so on. The privacy policy describes how the company treats sensitive information collected from users. The Platform for Private Preferences (P3P) [7] enables Web sites to express their privacy policies in a standard format that can be interpreted automatically by user agents. Thus, user agents can automate decision-making by comparing a company-defined privacy policy and user-specified privacy preferences. P3P, however, does not provide technical assurance that the sites will adhere to their respective privacy policies.

The Enterprise Privacy Authorization Language (EPAL) [8] provides fine-grained enterprise privacy policies, which employees within an organization are required to comply with. Whereas compliance to EPAL prevents the abuse of information by employees within an organization; it cannot, however, provide an assurance to users that the methods used by the organization to manage their personal information are secure.

Various researchers have attempted to provide users with the right of information access or services based on trustworthiness [9, 10]. Their researches were aimed to developing trust relationships among users. However, it is difficult to define a general method for developing trust relationships. This is because trustworthiness depends on user conditions and/or situations. Moreover, trust relationships are not directly connected to the prevention of information abuse.

Some researchers have proposed the use of mediator support to protect user information [11, 12]. Mediators could be considered as a cloud computing service. It might be difficult to prepare such a mediator because the mediator should be trusted by both the sides involved in the communication.

Chow et al [13] says that the lack of control in the cloud is a major concern and that sufficient transparency and auditability would be required in the operations of the cloud service provider. Wang et al [14] and Benaloh et al [15] propose a mechanism for storing encrypted data on the cloud while preventing the cloud service provider from reading the data, but allowing only those users who have the right of data access to retrieve the data. However, the goal of these researches is to prevent the access to data by cloud service providers; and not from those who have the right of data access.

Inada et al [16] and Miyamoto et al [17] propose methods to preserve privacy by forbidding the disclosure of the combination of information that may lead to the loss of anonymity. However, these methods cannot preserve privacy nor protect information when the loss of only a single piece of information (e.g., credit card number or password) cause a problem.

Encapsulated Mobile Agent-based Privacy Protection (EMAPP) [18] is a model in which a user's sensitive information is protected by not allowing service providers to access the information directly. In EMAPP, each user is assigned an *encapsulated space* that manages his/her information. A service provider has a *mobile agent* that checks the user's information. When a user requests a service, a mobile agent migrates into the user's encapsulated space, checks the user's information, and then sends only its result back to the service provider. Because sensitive information is not released outside the encapsulated space, the information is protected. However, it is matter to verify the security of the mobile agent because the mobile agent may be malicious. Moreover, a service provider may not be able to trust the result received from a mobile agent, because a user could potentially alter the result sent back by it.

Further, the proposed methods do not allow users to determine how their information is protected. However, he/she should also be able to determine the manner in which his/her information is protected, because the owner of information is a user. In this paper, we propose a framework that enables users to select the manner in which their information is protected.

3 User Customized Data Protection Framework

Typically, a service provider provides a service only to those users, who pos- sess specific information such as an ID-password pair, a name, and an address. Thus, a user is required to provide such information to the service provider. Subsequently, the user no longer has control over the information released to the service provider. We propose a framework that would enable users to cus- tomize the manner in which their information is protected; this framework will be provided as a kind of an SaaS.

3.1 Security as a Service

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [19]. It is a major concern for users that data and processes are entrusted to cloud service providers [20, 21].

Security as a Service (SaaS) is a type of cloud computing service. Such ser- vices are provided by McAfee [2], Panda Security [3], Yahoo! Japan [4], Rakuten Ichiba [5], etc. These companies provide various security services such as malware detection, URL filtering, and anonymous trading services, which work on the cloud instead of on client-installed software. Such services require users to entrust the management of their information to the cloud service providers. This is one of major concerns for users when using cloud services. If we can confirm and verify that data and processes in the cloud are secure, then such concerns could be deminished. However, this is very difficult to achieve because of reasons such as security, privacy, and company regulation.

In our framework, policies for the protection of sensitive information are pro- vided as a SaaS service. Policies are available in the form of text-based information and are open to the public. Because anyone can view and verify the policies, the one used frequently can be considered to be reliable. When a user wishes to protect his/her information in a particular manner, he/she can select a policy that defines such type of information protection. The manner defined in the policy selected by the user can be embedded into a program. The service provider makes use of the user's information through this program. Consequently, our framework allows users to select the manner to protect their sensitive information by using the appropriate policies.

3.2 An Overview of Our Framework

A service provider has a program for offering services after verifying the information provided by users. We call this program the *original program*. To protect user information from the service provider in a manner selected by the user, the part of the original program that processes user information is replaced with operations for the realization of the policy chosen by the user. We call this program the *customized program*. By allowing a service provider to use the user's sensitive information through a customized program, the user can protect his/her information in a manner selected by him/her. The overview of our framework is illustrated in Figure 1.

The framework consists of users, service providers, a rule repository, and program conversion services. A user has sensitive information and requests a particular service to a service provider. The service provider has an original program, which provides the service to the users after verifying the user information and the *usage policy*, which defines what information the service provider requires, the purpose for which the information is used, and what operations are applied to the information. A rule repository and program conversion services are on the cloud as an SaaS. A rule repository stores *protection policies*, which defines the type of information protection. Program conversion services provide a service for creating a customized program from an original program according to the protection policy.

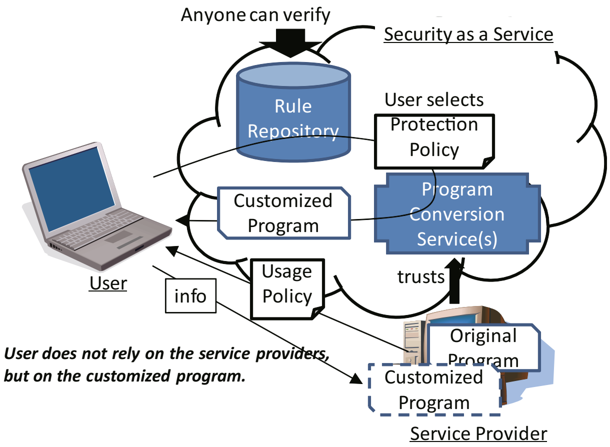


Fig. 1. An Overview of Our Proposed Framework

When a user uses a service, he/she receives a usage policy from the service provider. The usage policy describes the information required by the service provider, the purpose for which the information is used, and the operations applied to the information. Here, the operations applied to the user's information are fixed solely by the service provider. If the user does not trust the reliability of these operations, he/she selects a protection policy that satisfies the purpose of information use defined in the usage policy from the rule repository. Subsequently, the user asks a program conversion service to create a customized program according to his/her selected protection policy. By allowing the service provider to use his/her information through the customized program, the user can protect his/her information in a manner chosen by him/her. Thus, the user, along with the service provider, can take up the responsibility of his/her information protection.

3.3 Rule Repository

A rule repository stores protection policies that define the manner in which information is protected. Anyone can obtain any protection policy from the rule repository, and thus, anyone can verify any protection policy. Therefore, protection policies that are used frequently can be considered to be reliable. Therefore, users who have the ability to verify a protection policy can make sure that the protection policy surely protects information. However, almost users have no ability to verify a policy. For such a user, a rule repository should provide some criteria that users judge a protection policy reliable or not. The criteria would be possible, for example, to be given as a score calculated from a number of uses, troubles and the policy verified. Further, if the administrator of a rule repository has the burden to verify a protection policy when the policy is registered in, protection policies registered in would be considered to be reliable for users. Thus, almost users are not required to have no burden to verify a protection policy by themselves.

A protection policy is registered by volunteers. A volunteer could be a company or an individual. For example, a credit card company could register a protection policy for protecting a credit card number and recommends it to credit card users. Users can then use the recommended protection policy. In this case, users may not need to pay attention to whether the protection policy is reliable or not. Of course, if users are aware of a better policy to protect their information, they may/can choose that policy instead. In the case of using the recommended policy from the credit card company, the credit card company has to take the responsibility of the miuse case of a credit card number, while the user chooses a policy by himself, he has to take its responsibility.

A protection policy consists of an objective usage, the explanation of the policy, and the program conversion rules to realize the protection. Program conversion rules consist of the rules to convert an operation and a value of the original program to other operations and values, respectively. Thus, the program conversion rules enable the creation of a customized program from the original program.

3.4 Program Conversion Service

A program conversion service provides a service for creating a customized program from an original program according to a protection policy. Program conversion services are basically available on the cloud, and a large number of such program conversion services exist. The program conversion service is selected by the service providers. Because the program conversion service is selected by the service providers, only the most reliable program conversion service for the service providers will be selected. If there are no reliable program conversion services available to a service provider, the service provider can itself perform the task of a program conversion service. Thus, the customized program created by the program conversion service would be relied on by the service provider.

A user cannot, however, rely on the program conversion service, because it is selected by a service provider. Therefore, it is necessary for the user to verify the customized program is generated exactly as per the a protection policy selected by him/her.

Thus, we need 1) to show the service provider that the customized program was created by the program conversion service chosen by the service provider, and 2) to show the user that the customized program was exactly as per the pro- tection policy selected by the user. We solve the former issue by using the digital signature of the customized program. The later issue is solved by inspection of the customized program that is created according to the protection policy cho- sen by the user. The verification mechanism of a customized program is shown in Figure 2.

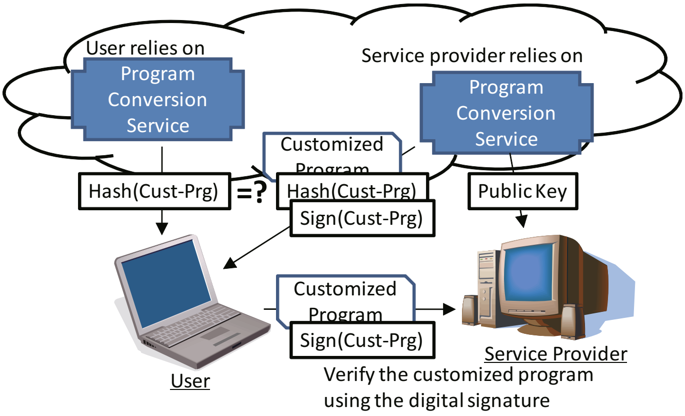


Fig. 2. Verification of a Customized Program

1. A service provider selects a program conversion service and receives a public key (of a public key cryptosystem) from it. Then, the service provider communicates the identifier of the program conversion service and the usage policy to the user.
2. Based on the usage policy, the user selects a protection policy from a rule repository, and asks the program conversion service chosen by the service provider to create a customized program according to the protection policy.
3. The program conversion service creates a customized program from the original program as per the protection policy. After that, it generates a digital signature and a hash value for the customized program, and sends them back to the user.

4. The user selects a reliable program conversion service and asks it to create a customized program according to the protection policy. (In this step, the user is also permitted to perform the task of the program conversion service, if there are no reliable program conversion services availabe to the user.)
5. The user receives the hash value of the customized program from the program conversion service chosen by the user. If its hash value is the same with the hash value mentioned in Steps 3, the user confirms that the customized program has been created as per the protection policy.
6. The user sends the customized program and its digital signature to the service provider. The service provider verifies the digital signature using the public key received in Step 1. If the digital signature is verified to be correct, it is proved that the the customized program has been created the program conversion service chosen by the service provider.

Thus, both the service provider and the user can confirm that the customized program is created exactly as per the protection policy.

operation	::=	`equal` | `greaterthan` | ...
description	::=	*how the protection policy protects information*
op-change	::=	`operation -> operation(s)`
var-change	::=	`var` *created from* `operation(var(s))` [*at user* | *service-provider*]
sendable	::=	`var(s)`, [*true* | *false*], [`condition`]

Fig. 3. Structure of Protection Policy

info	::=	*information the service provider requires*
operation	::=	`equal` | `greaterthan` | ...
var	::=	*arg used in the operation*
location	::=	(`var` | `operation`) [*received* | *created*] *at* `line-x`

Fig. 4. Structure of Usage Policy

3.5 Protection Policy

The customized program is created on the basis of the protection policy. The structure of a protection policy is shown in figure 3.

`Operation` specifies the targeted operation to which this protection policy can be applied. `Description` is written in natural language and describes the manner in which the protection policy protects user information. The protection manner written in `description` is defined using `op-change,` `var-change`, and `sendable`. `Op-change` is a rule for changing an operation to other operations to realize the protection written in `description`. For example, if `op-change` is ”f_a (x) -> f_b (y) && f_c (z)”, then f_a (x) is replaced by the combination of f_b (y) and f_c (z). `Var-change` is a rule for changing a value. For example, if `var-change` is `y` *created from* `hash(x)` *at user*, value `y` is created from value `x` using a hash function on the user-side. `Sendable` defines whether it is permitted to send `var` to other entities (who is a user/service-provider when `var` is created at the service-provider/user side) under `condition`.

A user understands a protection policy by viewing the text in `operation` and `description`, and asks a program conversion service to create a customized program based on it. Then, the program conversion service creates a customized program according to `op-change`, `var-change`, and `sendable`.

3.6 Usage Policy

A usage policy is prepared by the service provider, and it defines the purpose of information use. From the usage policy, the user can know how his/her informa- tion will be used. Figure 4 shows the structure of a usage policy.

A usage policy consists of `info`, `operation`, `var` and `location`. `Info` describes what information a service provider requires. `Operation` describes how the information is operated. A user uses `operation` and `info` to select a suitable protection policy for protecting `info` in `operation`. `Var` defines a variable used in `operation`. `Location` indicates where the operation and variables appear in the original program. A program conversion services can create a customized program by replacing the programs indicated in `location` according to a protection policy.

operation	= `equal`
description	= *Can protect arg0 by using hash-value, when user has arg0, and service-provider has arg1*
op-change	= `equal(arg0, arg1) -> equal(hash-pass, hash-y)`
var-change	= `r` *created from* `genRand()` *at user*
var-change	= `hash-pass` *created from* `hash(arg0, r)` *at user*
var-change	= `hash-y` *created from* `hash(arg1, r)` *at service-provider*
sendable	= (`arg0, arg1, hash-y`), *false*
sendable	= (`hash-pass, r`), *true*

Fig. 5. Protection Policy for Using Hashed Form Password

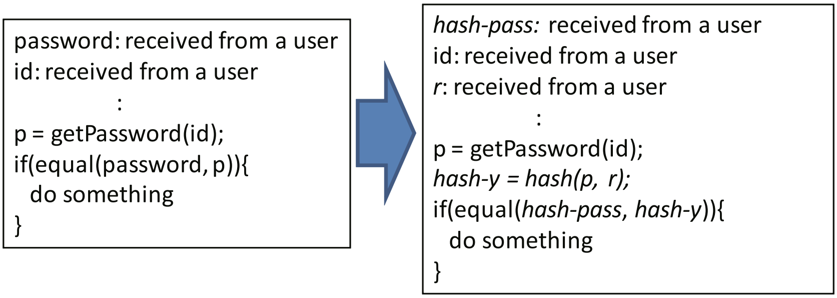
4 An Example Scenario

We describes an scenario that involves protecting a password from a service provider. In this scenario, a service provider requires a user's ID and password for user authentication. However, the user considers his/her password to be safe only when they are used in the hashed form, and a protection policy shown in Figure 5 is stored in a rule repository. Here, we should note that if the user believes another way to be safer, he/she will use that way for the protection of his/her password, even if it is vulnerable.

The service provider has an original program to check the user's ID and password. The original program would be as shown on the left of Figure 6. At first, the user receives a usage policy of the password and the identifier of a program conversion service from the service provider. In this case, the usage policy would be ``info=password, operation=equal, var=password, location=password received at line 1, equal at line 5, ...". The user selects a protection policy (shown in Figure 5) enabling his/her password to be used in the hashed form from a rule repository. Then, the user asks the program conversion service to create a customized program as per the protection policy. In this case, the protection policy would be as shown in Figure 5.

The creation of a customized program is done as follows. *Equal(password, p)* is replaced with *equal(hash-pass, hash-y)* according to `op-change`. `Var-change` describes the creation of a random number r, hash-pass from arg0 (which is the password possessed by the user), and hash-y from arg1 (which is the password stored by the service provider)[3] . In addition, because `sendables` of the hash-pass and r are true, hash-pass and r are sent from the user to the service provider. Finally, the customized program shown on the right of Figure 6 is created.

After that, as described in section 3.4, the program conversion service sends the customized program with its hash value and digital signature back to the user. The user confirms that the customized program is created exactly as per the protection policy. The service provider verifies the customized program using the digital signature. Finally, the service provider executes the customized program. Thus, the user can protect his/her password in a manner that he/she believes to be reliable, even if the service provider is involved in phishing or other malicious activities.



```
password: received from a user
id: received from a user
            :
p = getPassword(id);
if(equal(password, p)){
  do something
}
```

```
hash-pass: received from a user
id: received from a user
r: received from a user
            :
p = getPassword(id);
hash-y = hash(p, r);
if(equal(hash-pass, hash-y)){
  do something
}
```

Fig. 6. Left is an Original Program and Right is a Customized Program

5 Conclusion

Some service providers request users to provide sensitive information. Once a user releases sensitive information to a service provider, the user has no control over it.

[3] Practically r should be generated by the service provider in order to prevent a user from trying replay attacks.

The information would be under the control of the service provider. Thus, we must exercise caution while divulging sensitive information to service providers. To enable better protection of sensitive information, we have proposed a framework within which users can protect their sensitive information in a manner they believe to be reliable.

In our framework, rule repositories and program conversion services are of- fered as a *Security as a Service*. A user selects a protection policy that defines the type of information protection he/she desires from a rule repository. Then, the user asks a program conversion service to create a customized program that incorporates the desired information protection defined in the protection policy. Finally, by allowing the service provider to use his/her information through the customized program, the user can ensure that his/her information is protected in a manner chosen by him/her. Thus, the user, along with the service provider, can take up the responsibility of his/her information protection.

The future studies will include the implementation of our framework and evaluation of its efficacy. We also need to discuss in greater detail the structure of the usage policy and the protection policy.

Acknowledgments

We would like to thank Dr. Naohiko Uramoto for his valuable comments and advice. This research has been supported by a Grant-in-Aid for Young Scientists (B), 23700098.

References

1. Japan Network Security Association. Information Security Incident Survey Report ver.1.0, http://www.jnsa.org/result/index.html
2. McAfee Security-as-a-Service, http://www.mcafee.com/us/saas/index.html
3. Panda Cloud Protection, http://cloudprotection.pandasecurity.com/
4. Yahoo! Auction - Safety Payment Service, http://special.auctions.yahoo.co.jp/html/uketorigo/
5. Rakuten Safety Trading Services, http://event.rakuten.co.jp/anshin/
6. Stinson, D.R. (ed.): Cryptography: Theory and Practice, Crc Pr I Llc (1995)
7. P3P project, http://www.w3.org/P3P
8. The EPAL 1.1, http://www.zurich.ibm.com/security/enterpriseprivacy/epal/
9. Theodorakopoulos, G., Baras, J.: Trust Evaluation in Ad-Hoc Networks. In: WiSe 2004, pp. 1–10 (2004)
10. Xiu, D., Liu, Z.: Trust Model for Pervasive Computing Environments. In: FTDCS 2004, pp. 80–85 (2004)
11. Karabulut, Y.: Towards a Next-Generation Trust Management Infrastructure for Open Computing Systems. In: SPPC 2004 (2004)
12. Pearce, C., Bertok, P., Schyndel, R.: Protecting Consumer Data in Composite Web Services. In: IFIP/SEC 2005 (2005)

13. Chow, R., Golle, P., Jakobson, M., Shi, E., Staddon, J., Masuoka, R., Molina, J.: Controlling Data in the Cloud: Outsourcing Computation Without Outsourcing Control. In: CCSW 2010, pp. 85–90 (2009)
14. Wang, W., Li, Z., Owens, R., Bhargave, B.: Secure and Efficient Access to Outsourced Data. In: CCSW 2010, pp. 55–65 (2009)
15. Benaloh, J., Chase, M., Horvitz, E., Lauter, K.: Patient Controlled Encryption: En- suring Privacy of Electronic Medical Records. In: CCSW 2010, pp. 103–114 (2009)
16. Imada, M., Takasugi, K., Ohta, M., Koyanagi, K.: LooM: A Loosely Managed Privacy Protection Method for Ubiquitous Networking Environments. IEICE Trans. on Comm. J88-B(3), 563–573 (2005)
17. Miyamoto, T., Takeuchi, T., Okuda, T., Harumoto, K., Ariyoshi, Y., Shimojo, S.: A Proposal for Profile Control Mechanism Considering Privacy and Quality of Personalization Services. In: DEWS 2005, 6A-o1(2005)
18. Yamada, S., Kamioka, E.: Access Control for Security and Privacy in Ubiquitous Computing Environments. IEICE Trans. on Comm E88-B(3), 846–856
19. Mell, P., Grance, T.: The NIST Definition of Cloud Computing, `http://csrc.nist.gov/groups/SNS/cloud-computing/`
20. The result of questionnaire about Cloud Computing, `http://jp.fujitsu.com/about/journal/voice/enq/enq0905.shtml`
21. 2010 Analysis Report of the Market of Cloud Service in Japan, `http://www.idcjapan.co.jp/Press/Current/20100603Apr.html`

Centralizing Network Digital Evidences

Mohammed Abbas[1], Elfadil Sabeil[1], and Azizah Abdul Manaf[2]

[1] Faculty of Computer & Information Systems,
Universiti Teknologi Malaysia (UTM),
Johor Baru, Malaysia
mabbsaleh@gmail.com, alfadil.sabeel@yahoo.com
[2] Advanced Informatics School (AIS), Universiti Teknologi Malaysia (UTM),
Kuala Lumpur, Malaysia
azizah07@citycampus.utm.my

Abstract. The forensic community has long acknowledged only investigating operating system (computer) for the sake of discovering digital crimes secrets. However, these techniques are not reliable anymore in case when to be used to achieve investigation aims since the data of the operating system can be tampered with by an attacker himself. Hence, focusing on alternative fields; that is network forensic comes into picture. In this paper, a methodology to collect and centralize network digital evidences in order to come up with the reliable investigation is introduced. In a case study, the laboratory is designed and set up to examine the proposed solution toward network digital evidences and centralize them as well. Finally, the operating system forensic weaknesses are obviously proven, and then a successful solution to these shortcomings through collecting and centralizing network digital evidences to be used for the investigation is presented.

Keywords: Digital forensic, network forensic, digital crime investigation, computer forensic, malware, botnets.

1 Introduction

In general, traditional information technology environments consist of main critical digital components such as Routers, Firewalls, Intrusion Prevention Systems and operating systems used as servers in order to deliver its mission. Fig. 1 depicts an overview of these common parts of an environment that is available nowadays. Normally, these equipments being configured and assigned an Internet Protocol (IP) which explore and probes them all over the world means they could be accessible from outside to everyone. Actually, the mentioned feature presents risk toward an IT environment since it allows an attacker to bypass and circumvent the built security solutions in case there is a zero-day attack because everything is detectable and known form outside.

Recently, attackers have grown to be more intelligent due to investigations since they keep developing new techniques used to hide or overwrite the digital traces which might lead to grasp them. One of these expected crimes, overwriting all of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 310–320, 2011.

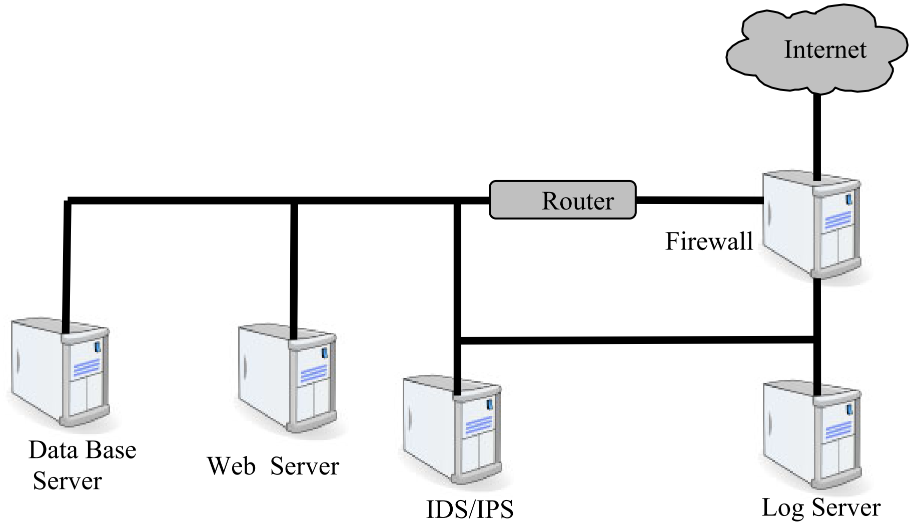


Fig. 1. Traditional IT Infrastructure

operating system digital traces, firewall logs files, or intrusion/Prevention Systems (IDS/IPS) logs files and so on. Furthermore, sometimes even worse, they use encrypted channels during conducting attacks which make digital traces analysis impossible without the decryption.

In the occurrence of attacks, it is enormously difficult to come up with a detailed analysis of how the attack happened and depicting what the steps were especially against skilled attackers who are clever enough at covering their tracks. The operating systems digital traces, Router logs files, Firewall logs files and intrusion detection alerts are unlikely to be sufficient for a serious investigation. Therefore, the efficient solution is in the area of Network Forensics; a dedicated investigation technology that allows for the capture, recording and analysis of network packets and events in order to conduct proper investigation [1]. In general, network forensics defined as is the capturing, recording, and analyzing of network packets and events for the sake of investigative purposes.

From results in this research, concentrating on network forensic is more accurate and much reliable since it allows setting up incorporated hidden nodes or hidden points in network environment that are not detectable by attackers to be used for capturing the desired suspected packets in investigation processes and then these packets should be centralized as well for the sake of simplifying investigations. Honeynet architecture is mainly used here to achieve research aim.

A Honeynet is an architecture which its purpose is basically to build a highly controlled network that control and capture all inbound and outbound network activities. Usually, within this architecture our Honeynets are placed. A Honeypot is a real system with valid services, open ports, applications and data files [2]. One of the key Honeynet architecture components is the Honeynet gateway which called Honeywall operating system. The Honeywall operating system is a very significant element in the Honeynet architecture since it captures and controls all of the inbound and outbound activities.

Centralizing collected digital evidences in an enterprise environment is an essential at various levels, one of the most important being that it allows security administrators and analyzers to monitor many systems in one central place. Once this digital evidences information is has centralized, it also allows them to achieve a more complete analysis and gives the ability to correlate events that have occurred. These centralized network's digital evidences can be used for networking investigation as alternative digital evidences [3][4][5].

2 Research Methodology

As Fig. 2 demonstrates, our centralizing network digital evidences methodology encompasses two logically dissimilar phases. (1) Digital evidences collection (capturing). (2) Digital evidences centralization.

Firstly, the goal of network digital evidences collection phase is simply to capture attackers' activities as many as possible. However, developing a robust scalable infrastructure to achieve this goal is challenging and is a target of numerous researchers [2][9]. In particular, any designed and developed infrastructure should scalable enough to support wide range of digital evidences collection. In addition, special actions must be implemented to prevent any part of system to behaving malfeasance. However, various different mechanisms are used in order to overcome the mentioned problems and therefore one solution (utility) is a candidate to represent each mechanism. In special, IPTable firewall utility represents firewall mechanism, Snort utility represents an intrusion prevention system [8], Logsys utility represents logging system [6] and lastly Sebek utility represents key logger for an encrypted network packets [3][4][5]. However, these utilities are explained in more details in next sections. Then in follows, the special infrastructure is shown.

An IPTables firewall installed on Honeywall OS basically to be used for capturing attackers' activities and actions against a victim since Honeywall is configured as a hidden media (in bridge mode) as depicted in Fig. 3.

Therefore, IPTables firewall is immune to be detected by attackers. By default, IPTables log its message to a /var/log/ directory. The next sample depicts an interested potential an attacker's activity he attempted as against a target:

```
Sep 7 08:56:25 mohd kernel: INBOUND TCP: In=br0 PHYSIN=eth0 OUT= br0

PHYSOUT= eth1 SRC=192.168.0.5 DST=10.0.0.7 LEN=64 TOS=0x00

PREC=0x00 TTL=48 ID=42222 DF PROTO=TCP SPT=1667 DPT=21

WINDOw=65535 RES=0x00 URGP=0
```

The previous firewall log output explains very significant information about an attacker's activity which to be used by investigators to reveal and analyze attacks and could be analyzed as following:

The Date and time: Sep 7 08:56:25

The Source IP address of an attacker: 192.168.0.5

The destination IP address of an attacker: 10.0.0.7

The protocol being used by an attacker: TCP

The source port of an attacker: 1667

The destination port of an attacker: 21

Snort with *Active Mode* or *Snort_InLine*, an intrusion prevention version of Snort; installed along with an IPTables firewall on Honewall in order to deliver the mission of the intrusion prevention system since highly required and could be used as another

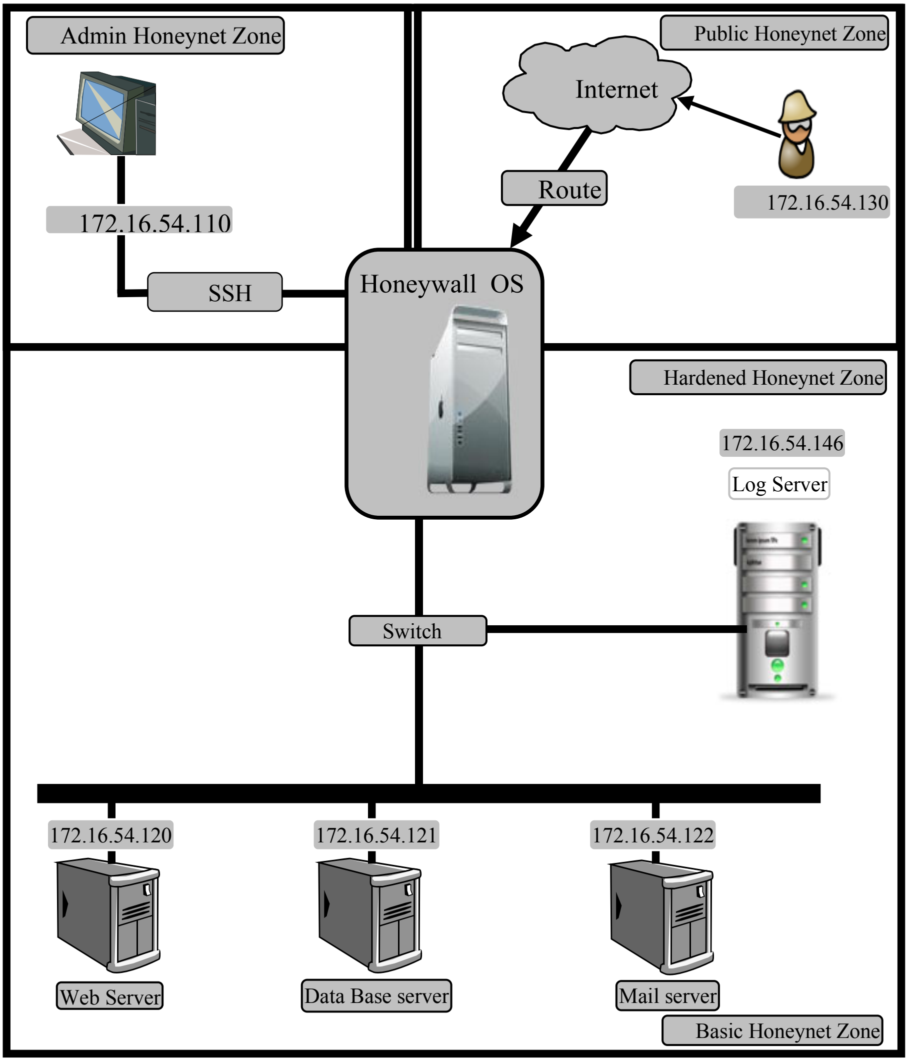


Fig. 2. Centralizing Network Digital Evidences

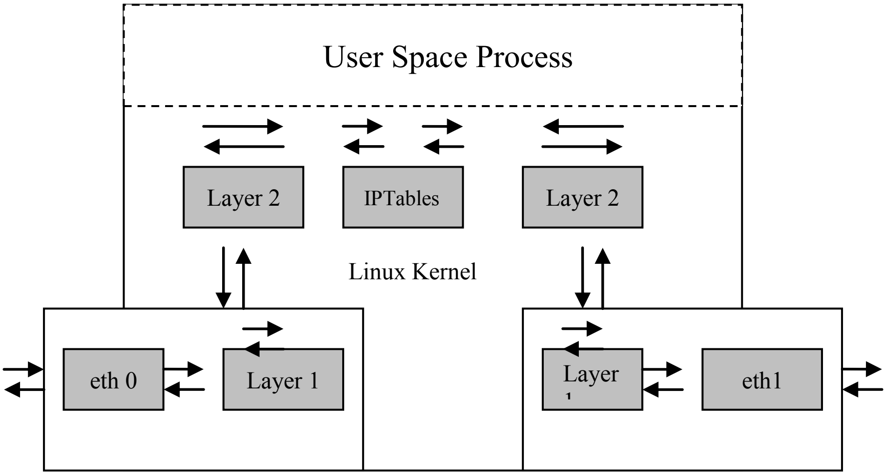


Fig. 3. IPTables and Bridging Mode

layer of protection and also digital evidences collection method as well. Fig. 4 shows the logical integration of Snort with IPTables on Honeywal OS.

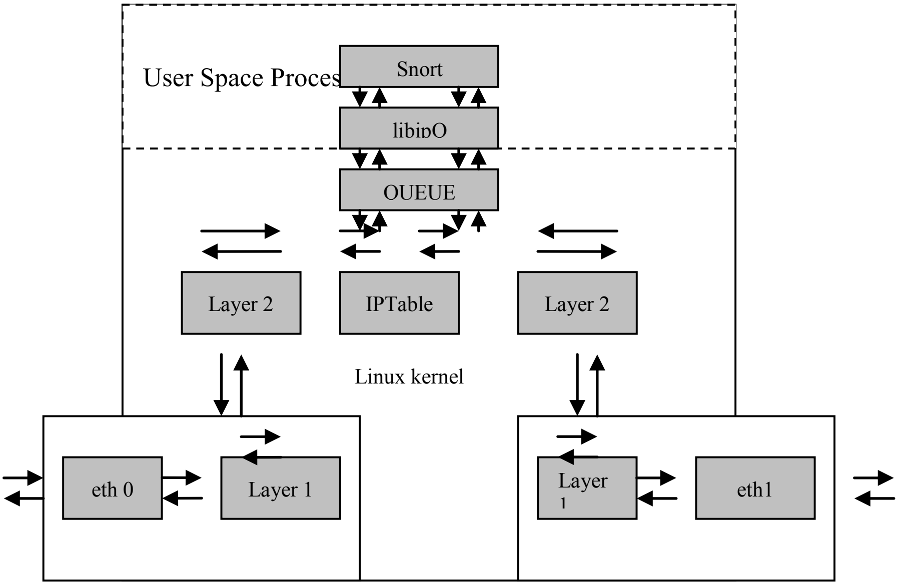


Fig. 4. Snort integrated with Honeywall OS

Snort mainly uses a rule driven language which combines benefits of signature, protocol, and anomaly based inspection methods [8]. Following next sample shows the collected digital evidences using this rule:

```
alert tcp any any -> any 80 (msg: "Sample alert"; classtype:misc-attack; sid: 2002973;
rev:1;)
          [**] [1:2002973:1] Sample alert [**]
          [Classification: Misc Attack] [Priority: 2]
          12/12-15:35:22.130162 test_client:35524 -> test_server:80
          TCP TTL:64 TOS:0x0 ID:35734 IpLen:20 DgmLen:52 DF
          ***A**** Seq: 0x5F3B46F0 Ack: 0x85067266 Win: 0xB7 TcpLen: 32
   TCP Options (3) => NOP NOP TS: 49925498 1529581
```

Log messages provide worth details about the events of devices and the applications running on these devices as well. Log messages could be used to discover and analyze security incidents, operational problems and policy violations which useful in auditing and forensics situations. The next sample shows collected digital evidence:

```
          mohd@ubuntu:~$ tail -f /var/log/messages
          Mar 27 11:34:35 ubuntu kernel: [ 165.459827] Bridge firewalling registered 44
          Mar 27 11:34:35 ubuntu kernel: [ 165.563138] Bluetooth: SCO socket layer
initialized
   Mar 27 11:34:39 ubuntu kernel: [ 170.004085] eth4: link up, 100Mbps, full-duplex
```

However, the outputs has generated by typing tail command. The analysis of The first output file analyzed as following:

```
          Time: Mar 27 11:34:35
          Host: ubuntu
          Facility: kernel
          Message: [ 165.459827] Bridge firewalling registered
```

However, in fact, the current iteration of Sebek tool is mainly designed not only to record keystrokes but to record all sys_read calls too. For an instance, if a file is has uploaded into Honeypot, Sebek immediately records the file which producing an identical copy. In general, Sebek consists of two components; a client and server as appeared in Fig. 5. The client part captures off Honeypot's data of a Honeypot and exports it to the server through the network. The server collects the data from one of two possible sources: the first is a live packet capture from the network, the second is a packet capture archive stored as a tcpdump formatted file. Once the data is collected then either will be uploaded into a relational database or the keystroke logs are immediately extracted. However, the client part resides entirely on kernel space within the Honeypot and implemented as a Loadable kernel Module (LKM). The client can record all data that a user accessed via the read() system call. This data is then exported to the server over the network in a manner that is difficult to detect from the Honeypot running Sebek. The server then received the data from all of the Honeypots sending data.

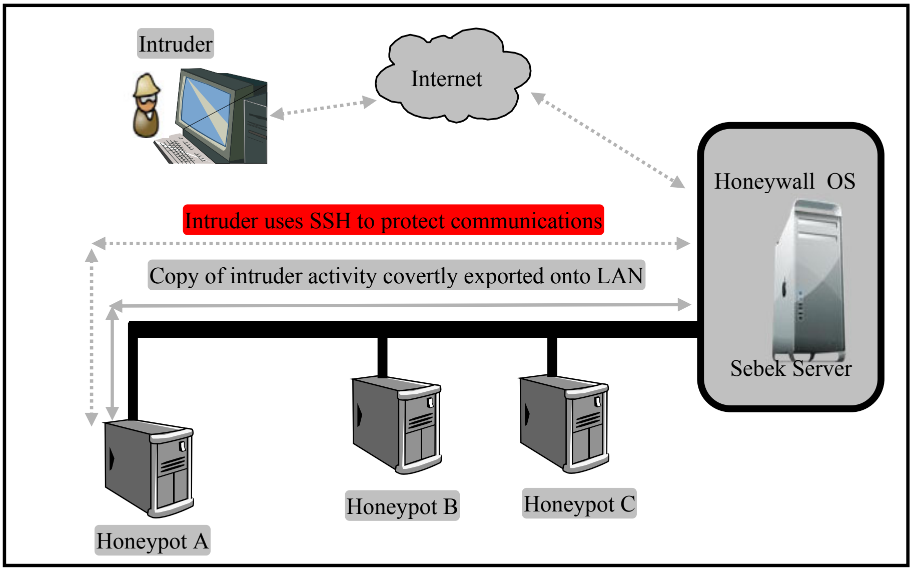

Intruder
Internet
Honeywall OS
Intruder uses SSH to protect communications
Copy of intruder activity covertly exported onto LAN
Sebek Server
Honeypot B
Honeypot C
Honeypot A

Fig. 5. Sebek Sever/Client Architecture

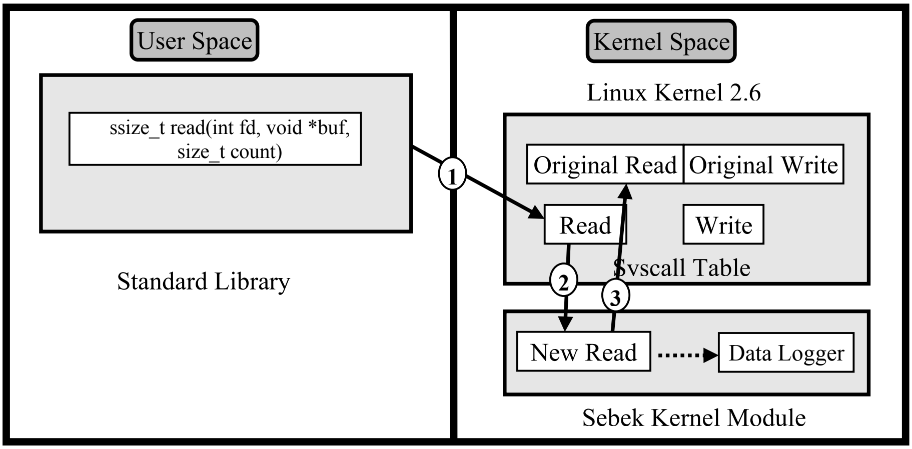


Fig. 6. Sebek Module Conceptual Logic

Moreover, Sebek generates its own packets and sends them directly to the device driver thus no ability for an attacker to block the packets or sniff them by using a network sniffer since Sebek uses its own implementation of the Raw Socket interface which programmed to silently ignore Sebek packets. This technique demonstrated clearly in Fig. 6. Sebek packets are defined as those that have both a predetermined destination UDP port and the proper magic number set in the Sebek header. If these two values match what is expected means this a packet to be ignored. The implementation simply does nothing with Sebek packets; it drops them on the floor and moves on to the next packet in the queue. The end result is that an intruder is unable to capture the Sebek packets.

Secondly, the main goal of network digital evidences centralization phase is merely to centralize the captured inbound and outbound network packets in arranged ways for simplifying investigation purposes since all evidences will be stored in one place. However, developing a robust solution to achieve this goal is challenging and should be scalable enough to support all sorts of collected evidences.. Therefore, all the previous evidence collection mechanisms are primarily used here and the next section demonstrates how they handled to achieve this aim.

The sequential logic of experiment's flow initially started by configuring and adapting main utilities that used in this research. The first step is to configure Log Server to accept and log the remote logs transferring that could be allowed by adopting the option to SYSLOGD=”-r”. Moreover, this configuration also involved for centralizing IPTables logs, since IPTables logs to system's logs. On the client side (Honeypots), their configuration option should be changed too to transfer their logs to the remote log server which achieved by adding **.* @172.16.54.146* into /etc/syslog.conf file. Then, Snort utility has adopted perfectly for capturing and recording an unencrypted channels properly. The captured network packets are recoded twice; into database and into flat files. Finally, Sebek server/client architecture has configured as well to deal with an encrypted channels. The Internet Protocol (IP=172.16.54.1) has assigned to Sebek Server which will be used to receive captured Sebek Clients packets. Also, UDP port number 1101, accept and log, and magic value options has chosen. After that, Sebek Client has installed into all honeypots and configured too. Sebek Client has adapted it's variables like UDP Port Number and Internet Protocol according to Sebek Server.

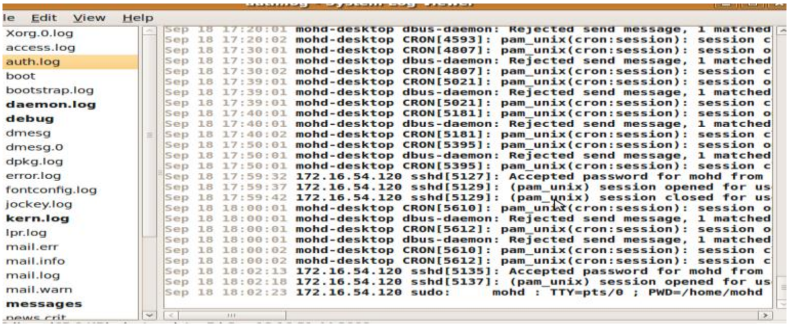
3 Results and Analysis

In fact, a case study has been used here to evaluate the designed methodology and test the proposed solution which mainly based on network digital evidences rather than computer digital evidences. A scenario constructed as following:

The production environment consists physically of four zones which each zone has its own components, tools and requirements. The first zone is a Basic Honeypots Zone which acts as a local network that held company's servers. These servers are Web Server and Secure Shell (IP=172.16.54.120), Data Base server, and Mail Server. Ubuntu Server operating system is basically installed and configured on all of these servers according to their requirements. Then, Sebek client utility also installed on Web Server in order to capture Web Server's activities that applied through a secure channel and then immediately transfer them to Sebek Server . After that, Honeywall OS is installed on a local network and acts as entry point of the lab's network. The Honeywall OS consists of various tools that used to collect and centralize the network's evidences. These tools are SNORT 2.8.4 application that used as main sniffer and also as Intrusion Detection or Prevention Systems (IDS/IPS). SNORT application is installed on Honeywall operating system and then customized in order in to log to a data base and dump network's packets. Lastly, Sebek Server is installed on Honeywall OS and configured to collect and log the encrypted connections to a data base. The second zone is a Hardened Honeypot Zone that configured just to allow connection from specified production environment's servers. This zone has log Server (IP=172.16.54.130) which mainly used to record all of production environment's servers activities. Indeed, Log Server access restrictions have hardened by using of an IPTables firewall. The third zone that is a Administrator Honeypot Zone which basically is used just by the administrator in order to configure and adapt Honeywall OS. The accessibility to a Honeywall OS only allowed through an encrypted channel by using of Secure Shell Server (IP=172.16.54.110). The fourth zone is a Public Internet Zone which is generally offers the production environment services to be accessible for the world. This zone is often untrusted and has users they might be legal or illegal.

However, an attacker compromised the production environment remotely when he came up with the root password through brute forcing SSH server. He then firstly uploaded two scripts files used for obfuscating crimes traces. These files mainly launched to remove chosen files and folders such as log's files and folders within web server. After that, he connected through an encrypted channels (SSH) to avoid intrusion detection systems [7]. Then, he uploaded another public key file to encrypt the web application and data base and overwrite the original files as well.

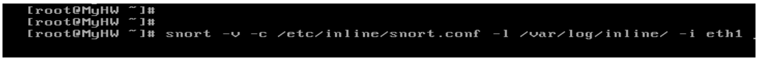
Now, analyzing results collected by the proposed solution helps us to discover digital crime's clues. Firstly, the attacker intentionally tried to login SSH server through brute force attack and he succeeds for logging as Fig. 7 depicts that.



```
Sep 18 17:20:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 17:20:02 mohd-desktop CRON[4593]: pam_unix(cron:session): session c
Sep 18 17:30:01 mohd-desktop CRON[4807]: pam_unix(cron:session): session o
Sep 18 17:30:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 17:30:02 mohd-desktop CRON[4807]: pam_unix(cron:session): session c
Sep 18 17:39:01 mohd-desktop CRON[5021]: pam_unix(cron:session): session o
Sep 18 17:39:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 17:39:01 mohd-desktop CRON[5021]: pam_unix(cron:session): session c
Sep 18 17:40:01 mohd-desktop CRON[5181]: pam_unix(cron:session): session o
Sep 18 17:40:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 17:40:02 mohd-desktop CRON[5181]: pam_unix(cron:session): session c
Sep 18 17:50:01 mohd-desktop CRON[5395]: pam_unix(cron:session): session o
Sep 18 17:50:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 17:50:01 mohd-desktop CRON[5395]: pam_unix(cron:session): session c
Sep 18 17:59:32 172.16.54.120 sshd[5127]: Accepted password for mohd from
Sep 18 17:59:37 172.16.54.120 sshd[5129]: (pam_unix) session opened for us
Sep 18 17:59:42 172.16.54.120 sshd[5129]: (pam_unix) session closed for us
Sep 18 18:00:01 mohd-desktop CRON[5610]: pam_unix(cron:session): session o
Sep 18 18:00:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 18:00:01 mohd-desktop CRON[5612]: pam_unix(cron:session): session o
Sep 18 18:00:01 mohd-desktop dbus-daemon: Rejected send message, 1 matched
Sep 18 18:00:02 mohd-desktop CRON[5610]: pam_unix(cron:session): session c
Sep 18 18:00:02 mohd-desktop CRON[5612]: pam_unix(cron:session): session c
Sep 18 18:02:13 172.16.54.120 sshd[5135]: Accepted password for mohd from
Sep 18 18:02:18 172.16.54.120 sshd[5137]: (pam_unix) session opened for us
Sep 18 18:02:23 172.16.54.120 sudo:     mohd : TTY=pts/0 ; PWD=/home/mohd
```

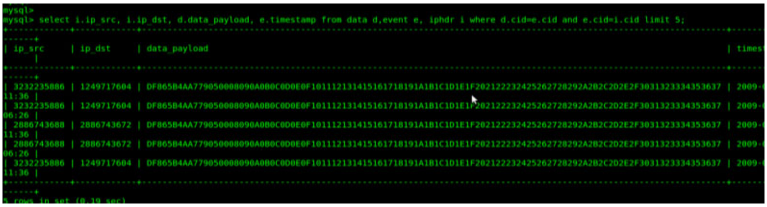
Fig. 7. Log Server received Honeypots Logs

Then, he uploaded two scripts namely BH-FE.rb [9] and BH-LSC.pl [10] through Secure Shell server (SSH) in order to bypasses the intrusion detection system (IDS). After that, he encrypted the web application files and the data base files too by using an uploaded public key. Finally and intentionally launched BH-FE.rb [9] and BH-LSC.pl [10] scripts in order to destroy the log files which might used for forensic investigation. All of an attacker's commands, and encrypted or unencrypted operations captured by Snort and Sebek server and then logged into data base. The following figures show these results:



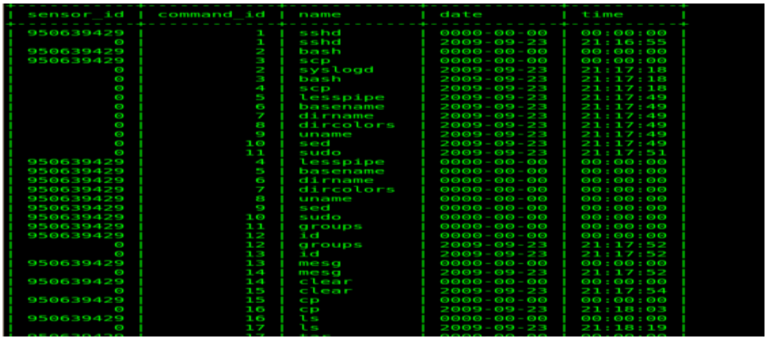
```
[root@MyHW ~]#
[root@MyHW ~]#
[root@MyHW ~]# snort -v -c /etc/inline/snort.conf -l /var/log/inline/ -i eth1
```

Fig. 8. Launching Snort Utility



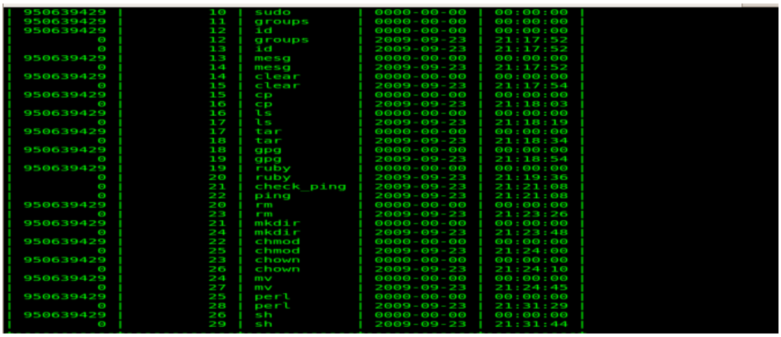
```
mysql>
mysql> select i.ip_src, i.ip_dst, d.data_payload, e.timestamp from data d,event e, iphdr i where d.cid=e.cid and e.cid=i.cid limit 5;
+------------+------------+------------------------------------------------------------------------------------------------------+-------
-------+
| ip_src     | ip_dst     | data_payload                                                                                         | times
       |
+------------+------------+------------------------------------------------------------------------------------------------------+-------
-------+
| 3232235886 | 1249717604 | DF865B4AA779050008090A0B0C0D0E0F101112131415161718191A1B1C1D1E1F202122232425262728292A2B2C2D2E2F3031323334353637 | 2009-
11:36 |
| 3232235886 | 1249717604 | DF865B4AA779050008090A0B0C0D0E0F101112131415161718191A1B1C1D1E1F202122232425262728292A2B2C2D2E2F3031323334353637 | 2009-
06:26 |
| 2886743688 | 2886743672 | DF865B4AA779050008090A0B0C0D0E0F101112131415161718191A1B1C1D1E1F202122232425262728292A2B2C2D2E2F3031323334353637 | 2009-
11:36 |
| 2886743688 | 2886743672 | DF865B4AA779050008090A0B0C0D0E0F101112131415161718191A1B1C1D1E1F202122232425262728292A2B2C2D2E2F3031323334353637 | 2009-
06:26 |
| 3232235886 | 1249717604 | DF865B4AA779050008090A0B0C0D0E0F101112131415161718191A1B1C1D1E1F202122232425262728292A2B2C2D2E2F3031323334353637 | 2009-
11:36 |
+------------+------------+------------------------------------------------------------------------------------------------------+-------
-------+
5 rows in set (0.19 sec)
```

Fig. 9. Network Evidences Logged by Snort



```
+-------------+------------+--------------+------------+----------+
| sensor_id   | command_id | name         | date       | time     |
+-------------+------------+--------------+------------+----------+
|   950639429 |          1 | sshd         | 0000-00-00 | 00:00:00 |
|           0 |          1 | sshd         | 2009-09-23 | 21:16:55 |
|   950639429 |          2 | bash         | 0000-00-00 | 00:00:00 |
|   950639429 |          3 | scp          | 0000-00-00 | 00:00:00 |
|           0 |          2 | syslogd      | 2009-09-23 | 21:17:18 |
|           0 |          3 | bash         | 2009-09-23 | 21:17:18 |
|           0 |          4 | scp          | 2009-09-23 | 21:17:18 |
|           0 |          5 | lesspipe     | 2009-09-23 | 21:17:49 |
|           0 |          6 | basename     | 2009-09-23 | 21:17:49 |
|           0 |          7 | dirname      | 2009-09-23 | 21:17:49 |
|           0 |          8 | dircolors    | 2009-09-23 | 21:17:49 |
|           0 |          9 | uname        | 2009-09-23 | 21:17:49 |
|           0 |         10 | sed          | 2009-09-23 | 21:17:49 |
|           0 |         11 | sudo         | 2009-09-23 | 21:17:51 |
|   950639429 |          4 | lesspipe     | 0000-00-00 | 00:00:00 |
|   950639429 |          5 | basename     | 0000-00-00 | 00:00:00 |
|   950639429 |          6 | dirname      | 0000-00-00 | 00:00:00 |
|   950639429 |          7 | dircolors    | 0000-00-00 | 00:00:00 |
|   950639429 |          8 | uname        | 0000-00-00 | 00:00:00 |
|   950639429 |          9 | sed          | 0000-00-00 | 00:00:00 |
|   950639429 |         10 | sudo         | 0000-00-00 | 00:00:00 |
|   950639429 |         11 | groups       | 0000-00-00 | 00:00:00 |
|   950639429 |         12 | id           | 0000-00-00 | 00:00:00 |
|           0 |         12 | groups       | 2009-09-23 | 21:17:52 |
|           0 |         13 | id           | 2009-09-23 | 21:17:52 |
|   950639429 |         13 | mesg         | 0000-00-00 | 00:00:00 |
|           0 |         14 | mesg         | 2009-09-23 | 21:17:52 |
|   950639429 |         14 | clear        | 0000-00-00 | 00:00:00 |
|           0 |         15 | clear        | 2009-09-23 | 21:17:54 |
|   950639429 |         15 | cp           | 0000-00-00 | 00:00:00 |
|           0 |         16 | cp           | 2009-09-23 | 21:18:03 |
|   950639429 |         16 | ls           | 0000-00-00 | 00:00:00 |
|           0 |         17 | ls           | 2009-09-23 | 21:18:19 |
```

Fig. 10. Network Evidences Logged by Sebek Architecture



```
|   950639429 |          10 | sudo         | 0000-00-00 | 00:00:00 |
|   950639429 |          11 | groups       | 0000-00-00 | 00:00:00 |
|   950639429 |          12 | id           | 0000-00-00 | 00:00:00 |
|           0 |          12 | groups       | 2009-09-23 | 21:17:52 |
|           0 |          13 | id           | 2009-09-23 | 21:17:52 |
|   950639429 |          13 | mesg         | 0000-00-00 | 00:00:00 |
|           0 |          14 | mesg         | 2009-09-23 | 21:17:52 |
|   950639429 |          14 | clear        | 0000-00-00 | 00:00:00 |
|           0 |          15 | clear        | 2009-09-23 | 21:17:54 |
|   950639429 |          15 | cp           | 0000-00-00 | 00:00:00 |
|           0 |          16 | cp           | 2009-09-23 | 21:18:03 |
|   950639429 |          16 | ls           | 0000-00-00 | 00:00:00 |
|           0 |          17 | ls           | 2009-09-23 | 21:18:19 |
|   950639429 |          17 | tar          | 0000-00-00 | 00:00:00 |
|           0 |          18 | tar          | 2009-09-23 | 21:18:34 |
|   950639429 |          18 | gpg          | 0000-00-00 | 00:00:00 |
|           0 |          19 | gpg          | 2009-09-23 | 21:18:54 |
|   950639429 |          19 | ruby         | 0000-00-00 | 00:00:00 |
|           0 |          20 | ruby         | 2009-09-23 | 21:19:36 |
|           0 |          21 | check_ping   | 2009-09-23 | 21:21:08 |
|           0 |          22 | ping         | 2009-09-23 | 21:21:08 |
|   950639429 |          20 | rm           | 0000-00-00 | 00:00:00 |
|           0 |          23 | rm           | 2009-09-23 | 21:23:26 |
|   950639429 |          21 | mkdir        | 0000-00-00 | 00:00:00 |
|           0 |          24 | mkdir        | 2009-09-23 | 21:23:48 |
|   950639429 |          22 | chmod        | 0000-00-00 | 00:00:00 |
|           0 |          25 | chmod        | 2009-09-23 | 21:24:00 |
|   950639429 |          23 | chown        | 0000-00-00 | 00:00:00 |
|           0 |          26 | chown        | 2009-09-23 | 21:24:10 |
|   950639429 |          24 | mv           | 0000-00-00 | 00:00:00 |
|           0 |          27 | mv           | 2009-09-23 | 21:24:45 |
|   950639429 |          25 | perl         | 0000-00-00 | 00:00:00 |
|           0 |          28 | perl         | 2009-09-23 | 21:31:29 |
|   950639429 |          26 | sh           | 0000-00-00 | 00:00:00 |
|           0 |          29 | sh           | 2009-09-23 | 21:31:44 |
```

Fig. 11. Network Evidences Logged by Sebek Architecture (Cont.)

4 Conclusion

In this research, the importance of network digital evidences regarding to digital crimes investigations rather than computer digital evidences (since they could be overwritten after conducting attacks) is deeply discussed and stated. The magical key for overcoming computer digital evidences weaknesses is to hide intermediate media in stealth mode between production environment and the outside world for capturing all or desired network inbound and outbound activities and it should not detectable by anyone. In fact, Honeynet Architecture used here to present this media since it's main component, Honeywall OS configured on bridged mode to achieve research's aim. Then, however, various utilities like IPTables, Snort, Sebek and Log server have adopted perfectly for the sake of collecting and centralizing network digital evidences and then how use these evidences for investigating digital crimes.

References

1. Almulhem, A., Issa Traore, I.: Experience with Engineering a Network
2. Spitzner, L.: Honeypots:Tracking Hackers. Addison-Wesley, Reading
3. Honeynet group. Know your Enemy, 2nd edn. Addison-Wesley, Reading
4. Honeynet Project. A kernel based data capture tool. Honeynet Project, 1–21
5. Honeynet group. Know your Enemy, 1st edn. Addison-Wesley, Reading
6. BalaBit, Distributed syslog architectures with syslog-ng Premium edn. BalaBit IT Security, pp. 1–12 (2007)
7. Heather, M.L. S.: Intrusion Detection. SANS Institute, 2–6
8. Ramirez, G., Caswell, B., Rathuas, N.: Nessus, Snort and Ethereal. Syngress Publishing, Inc., Rockland (2005)
9. BH-FE.rb script, http://aalagha.com/blog/2008/09/09/bh-final-eraser-version-05
10. BH-LSC.pl script, http://aalagha.com/blog/2008/04/20/bhlsc-linux-servercleaner

Challenges and Opportunities in the Information Systems Security Evaluation and Position of ISO / IEC 15408

Nasser Vali and Nasser Modiri

Tehran North Branch Azad University, Tehran, Iran
Zanjan Azad University, Zanjan, Iran
{Naser_vali,NasserModiri}@yahoo.com

Abstract. Organizations would encounter with challenges which leaving them would be impossible without any systematic and engineering approach and without any preparation of Secure Information System. The most important and greatest challenge is related to security of area that provides Information Systems. The main contribution of this work is providing a security standard-based process for software product line development. It is based on categories vulnerabilities and some concept of software engineering and use of the redefinition of information system life cycle, which integrated by Common Criteria (ISO/IEC 15408) controls into the product line lifecycle. Present approach reduces the complexity and ambiguity inherent in the information systems security in the engineering, well-defined, repeatability process.

Thus, the security organizations which implement secure products ensure the security level their product and use time-cost effective and engineering process to improve their future product.

Keywords: Information System (IS), Security requirement, Security Strategies, Security Engineering, Security Evaluation, Security Policy, ISO/IEC 15408.

1 Presentation

No country would have any kinds of development except if its people has reached to the high level of maturity in term of information and this would be obtained only by increasing information and knowledge day to day. The role of Information system (IS) is obvious in countries toward development working as_a tool for information sharing and universal access. And scientists of_that community are responsible for its development and expend it. Entering to this area without any systematic approach and preparation of IT and its architecture beside communication technology would encounter the country with challenges which leaving it would be impossible. The most important and greatest challenge is related to security of area that provides technologies. The area of security information exchange of any country is depending on several factors which needs different measures at the national level. Fortunately there are many security standards in this field like: ISO/IEC 27000i, ISO/IEC 17799, ISO/IEC 13355 and ISO/IEC 15408…, recently special attention have been paid on this subject.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 321–334, 2011.

2 Introduction

Information is considered as vital lifeline in organizations and developed institution and scientific societies, in the way towards a modern IT organization based on IT necessary measures should be considered in regarding the protection of Information system (IS). Success in security information depends on protection of IS against attacks, so it can be select strategically based on 3 elements:

Detection, Protection and Response. Reliable protection, on time detection and appropriate response, are the cases that a security system has to respect them as shows in fig1.

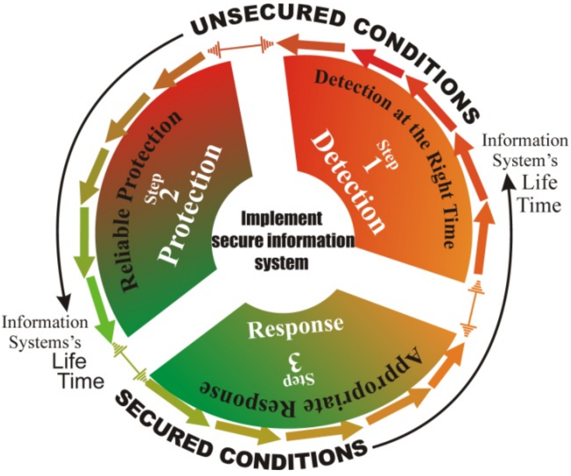

UNSECURED CONDITIONS
Detection at the Right Time
Step 1
Detection
Information System's Life Time
Reliable Protection
Step 2
Protection
Implement secure information system
Response
Step 3
Appropriate Response
Information Systems's Life Time
SECURED CONDITIONS

Fig. 1. Security Strategy Steps

First step, Detection; has shown in red, meant to unsafe (insecure) conditions in the system. In this time we needed to accurately identifying possible attacks, threats and their risks.

Second step, Protection; meant toward securing and decisions about security policy and designing of organization security plans.

Third step, Response to; would fix security problems by implementation of principles and obligate to the policy to correct the information system.

The most important idea in this figure is animating and rotates around the product life cycle that each information system can encounter with new types of security problems around its life cycle.

Evaluation process in the ISO/IEC 15408, defines degree of confidentiality of security function of product and systems, and ensures the amount required for application of these requirements specifies. Evaluation results can help customers determine whether their product or systems with information system security is enough or whether the security risks are acceptable or not [1].

So we would have dynamic strategy to setup the security can be cause of durable system and enhance the level of security.

3 Security versus Safety

3.1 Comparison

We have to considered that two entirely were separate subject and one of them has their own place in IS, safety engineering is defined as structured and clear repeatable process to reduce the risk of unintentional damage in cases where un authorization activities by over valuable asset. Damage in this sphere by investors is understandable and tolerable.

Security engineering defined as Structure-oriented and methodical and repeatable and enlightened process about deliberate damage from unauthorized activities and trying to reduce this risk for investors which is intolerable an indiscoverable[5],[7].

Table 1. Some difference in safety engineering and security engineering

Security Engineering	Safety Engineering
Intentional damage and subversive	Unintentional damage and minor mistakes or errors
Challenges must be controlled and attacks are purposeful and aggressive	Challenges are less important and attacks are non-purposeful
Events meant to abuse the system	Events are considered awful for system
There are serious threat and damage and Deficiency	There are risk (probability of damage)
Structure-oriented and methodical and repeatable and enlightened process	as structured and clearly repeatable process

3.2 What Is the Information Systems Security

In fact, information systems security include protection of IS against a wide range of threats, with the aim at ensuring the continuity of business activities, minimizing risks of working and maximize return of investment and opportunities.

Information security includes confidentiality, integrity and availability of information and also included others such as authenticity, accountability, non-repudiation and reliability. In this few sentences we can understand that information systems security is an ambiguous and very complicated concept in information systems development [2], [3], [4]. At the first we should define some words that help us to realize Information Systems Security:

Security policy: We titled security policy as a giving support to management and IS security based on work requirement and relevant rules and regulations in a documentary collection, management has to prove the business objective based on a specific policy, support and commitment, IS security through the publication of this document and adherence to comprehensive IS security in organization.

Documentation of information system security policy: Document of ISSP must be approved by management; publishing to all staff and relevant external organizational groups is communicated.

Review of information security policy: Document of ISSP should be scheduled at intervals to be revised if significant changes are made and ensure of maintaining competence, adequacy and its effectiveness.

4 Challenges

4.1 Complexity of the Security Concept

Obviously to define this concept, we have to discuss two aspects of this concept separately. First, ”WHAT”: what is really needed, as stated objectives or evaluate and provide really clear what the more intellectual for people who involved in the zone of security evaluation.

Second, “HOW”: determine policy and create tools that automate security measures to introduce the evaluation the current process-oriented development, Structure-oriented and methodical and repeatable and enlightened [2].

4.2 Challenges of Information System Evaluation and Product Warranty

Evaluation process documentary has to be considered during the life cycle of the IS, this requires security process management model and a methodical, reproducibility and self optimum engineering process.

Estimating profit from the amount invested in the secure implementation process and security assurance process is very necessary and critical. But unfortunately it is very hard and ambiguous to perform criteria and guarantee the return of investment (ROI). It should be mentioned that the insurance security is not meaning higher percentage return of investment (ROI), and sometimes only trying to guarantee or security assurance will raise the production cost [2]!

Security evaluation is done on documentation in product lifecycle, not in product, and it shows integration of software engineering topics and security engineering and higher level of assurance means perform and realize more and more rules and criteria of engineering software[2],[5].

4.3 Vulnerabilities and Defects[1]

Steps should be taken to prevent vulnerabilities arising in IT products. To the extent feasible, vulnerabilities should be:

a) Eliminated
b) Minimized
c) Monitored

4.3.1 Cause of Vulnerabilities

Vulnerabilities can arise through failures in:

a) Requirements -- that is, an IT product may possess all the functions and features required of it and still contain vulnerabilities that render it unsuitable or ineffective with respect to security;

[1] Common criteria, ISO/IEC 15408, part: 3; ver3.1:2009

b) Development -- that is, an IT product does not meet its specifications and/or vulnerabilities have been introduced as a result of poor development standards or incorrect design choices;

c) Operation -- that is, an IT product has been constructed correctly to a correct specification but vulnerabilities have been introduced as a result of inadequate controls upon the operation [1].

4.3.2 Type of Vulnerability

In any infrastructure, which includes software, there are vulnerabilities that cannot be designed or built out in an economic sense and can only be mitigated against. These are intrinsic vulnerabilities. By contrast, there are vulnerabilities that can be designed and built out 'non-intrinsic ones' and therefore doing so removes the need for mitigation. Understanding the difference between intrinsic and non-intrinsic vulnerabilities is at the heart of good software assurance [8].

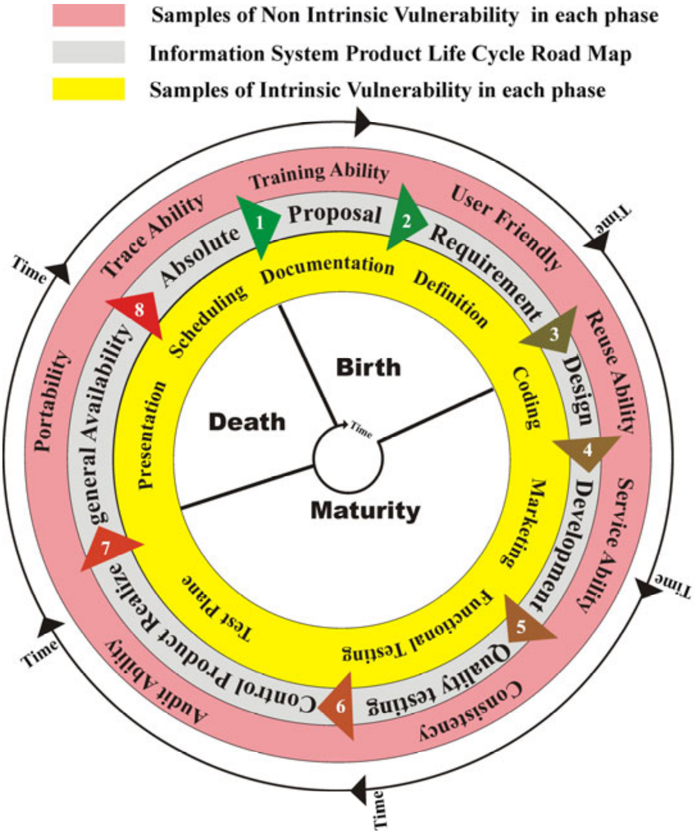


Fig. 2. Sample of intrinsic and non-intrinsic vulnerability in each step of Product Life Cycle

As you see in figure 2, we have defined 3 steps for product life cycle (PLC) as stages birth, maturity and death. Here we define Software life cycle (SLC) in 8 phases which has been similar to 3 steps of PLC and be able to classify every possible intrinsic and non-intrinsic vulnerabilities during the difference phase.

According to this figure proposal and requirement analysis phases and design production some extend phase are the parts of birth product step. Growth and Maturity stage of product including design, development, quality test, product control knowledge (limited release) phases and the product will reach to peak maturity when the phase of public release is starting and then must think of replacing products based on new requirement and passing the last shortcomings and vulnerabilities. This stage is product death which includes public access and absolute phases.

A solution for a non-intrinsic vulnerability might be the requirement to develop an authentication mechanism to protect against unauthorized use. By comparison, a solution for an intrinsic vulnerability might be installation of the system and its data on an isolated machine that has no internet connectivity. It is useful to note that the first is an example of a functional requirement, while the second is an example of a non-functional requirement [8]. Software assurance under ISO/IEC 15408 addresses both.

Reconciling this back to the original definition, software assurance is "a level of confidence that software is free from vulnerabilities, either intentionally designed into the software or accidentally inserted at any time during its life cycle." This part of the definition addresses the non-intrinsic vulnerabilities and the non-functional requirements caused by software flaws, whereas the rest of the definition, "and that the software functions in the intended manner," addresses the intrinsic vulnerabilities and the functional requirements for security [8].

5 Opportunities

5.1 Position of ISO/IEC 15408

5.1.1 CC Philosophy

The CC philosophy is that the threats to security and organizational security policy commitments should be clearly articulated and the proposed security measures should be demonstrably sufficient for their intended purposes.

Furthermore, those measures should be adopted that reduce the likelihood of vulnerabilities, the ability to exercise (i.e. intentionally exploit or unintentionally trigger) a vulnerability, and the extent of the damage that could occur from a vulnerability being exercised. Additionally, measures should be adopted that facilitate the subsequent identification of vulnerability and the elimination, mitigation, and/or notification that vulnerability has been exploited or triggered [1].

5.1.2 Application Domain

The CC does not contain security evaluation criteria pertaining to administrative security measures not related directly to the IT security functionality.

The evaluation of some technical physical aspects of IT security such as electromagnetic emanation control is not specifically covered, although many of the concepts will be applicable to that area.

The CC does not address the evaluation methodology under which the criteria should be applied.

The CC does not address the administrative and legal framework under which the criteria may be applied by evaluation authorities.

The procedures for use of evaluation results in accreditation are outside of scope of the CC.

The subject of criteria for the assessment of the inherent qualities of cryptographic algorithms is not covered in the CC.

To show requirements and ensure IT security operations, under the standard ISO / IEC 15408 the following two concepts are used:

Protection Profile infrastructure (PP)

The PP allows collecting and implementation completeness and reusability security requirement. PP can be use by customer for detecting and realizing secure product which meets their needs.

Security Target infrastructure (ST)

The ST shows security requirement and secure operation for evaluation system or special product which is called TOE (Target of Evaluation), ST is a base for evaluation according to the standard ISO/IEC 15408 and use with who evaluate on TOE.

The main concept of protection profiles (PP), packages of security requirements and the topic of conformance are specified and the consequences of evaluation, evaluation results are described. This part of the CC gives guidelines for the specification of Security Targets (ST) and provides a description of the organization of components throughout the model.

5.2 Functional Requirements Paradigm

TOE evaluation is concerned primarily with ensuring that a defined set of security functional requirements (SFRs) is enforced over the TOE resources. The SFRs define the rules by which the TOE governs access to and use of its resources, and thus information and services controlled by the TOE.

The SFRs may define multiple Security Function Policies (SFPs) to represent the rules that the TOE must enforce. Each such SFP must specify its scope of control, by defining the subjects, objects, resources or information, and operations which it applies. All SFPs are implemented by the TSF (see below), whose mechanisms enforce the rules defined in the SFRs and provide necessary capabilities. Those portions of a TOE that must be relied on for the correct enforcement of the SFRs are collectively referred as the TOE Security Functionality (TSF). The TSF consists of all hardware, software, and firmware of a TOE that is either directly or indirectly relied upon for security enforcement. The TOE may be a monolithic product containing hardware, firmware, and software [1]. Look figure 4 for more realization:

Alternatively a TOE may be a distributed product that consists internally of multiple separated parts. Each of these parts of the TOE provides a particular service for the TOE, and is connected to the other parts of the TOE through an internal

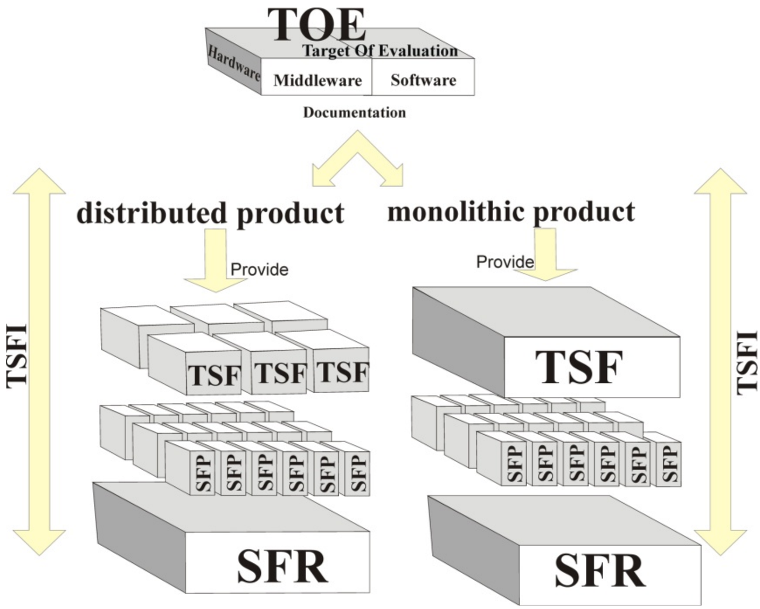


Fig. 3. The CC Functional Requirements Paradigm

communication channel. This channel can be as small as a processor bus, or may encompass a network internal to the TOE. The set of interfaces, whether interactive (man-machine interface) or programmatic (application programming interface), through which resources are accessed that are mediated by the TSF, or information is obtained from the TSF, is referred to as the TSF Interface (TSFI). The TSFI defines the boundaries of the TOE functionality that provide for the enforcement of the SFRs.

5.3 Security Function Components

This chapter defines the content and presentation of the functional requirements of the CC, in this paper we present only 3 components for more information see ISO/IEC 15408: part 2.

Class FAU: Security AUDIT

Security auditing involves recognizing, recording, storing, and analyzing information related to security relevant activities,

The resulting audit records can be examined to determine which security relevant activities took place and whom (which user) is responsible for them.

Class FCO: COMMINICATION

This class provides two families specifically concerned with assuring the identity of a party participating in a data exchange. These families are related to assuring the identity of the originator of transmitted information (proof of origin) and assuring the identity of the recipient of transmitted information (proof of receipt). These families ensure that an originator cannot deny having sent the message, nor can the recipient deny having received it.

Class FCS: PTOGRAPHIC SUPPORT

The TSF may employ cryptographic functionality to help satisfy several high-level security objectives. These include (but are not limited to): identification and authentication, non-repudiation, trusted path, trusted channel and data separation. This class is used when the TOE implements cryptographic functions, the implementation of which could be in hardware, firmware and/or software.

It is necessary for attention that defined component in cc, is not a final solution for solve all security problems, it is just a set of appropriate security requirement which help us to implement secure product that will meet costumer's requirement and just suggest them as common criteria. This security component according to current situation evaluation shows appropriate security requirements.

5.4 CC Assurance[2]

Assurance is grounds for confidence that an IT product meets its security objectives. Assurance can be derived from reference to sources such as unsubstantiated assertions, prior relevant experience, or specific experience. However, the CC provides assurance through active investigation. Active investigation is an evaluation of the IT product in order to determine its security properties.

5.4.1 The CC Evaluation Assurance Scale

The CC philosophy asserts that greater assurance results from the application of greater evaluation effort, and that the goal is to apply the minimum effort required to provide the necessary level of assurance. The increasing level of effort is based upon: Scope, Depth, Rigour.

5.4.2 Evaluation Assurance Level 1 (EAL1) -Functionally Tested

EAL1 is applicable where some confidence in correct operation is required, but the threats to security are not viewed as serious.EAL1 provides a basic level of assurance by a limited security target and an analysis of the SFRs in that ST using a functional and interface specification and guidance documentation, to understand the security behavior. The analysis is supported by a search for potential vulnerabilities in the public domain and independent testing (functional and penetration) of the TSF. EAL1 also provides assurance through unique identification of the TOE and of the relevant evaluation documents. This EAL provides a meaningful increase in assurance over unevaluated IT.

5.4.3 Evaluation Assurance Level 2 (EAL2) - Structurally Tested

EAL2 requires the co-operation of the developer in terms of the delivery of design information and test results, but should not demand more effort on the part of the developer than is consistent with good commercial practice. As such it should not require a substantially increased investment of cost or time. EAL2 is therefore applicable in those circumstances where developers or users require a low to moderate level of independently assured security in the absence of ready availability of the complete development record. Such a situation may arise when securing legacy systems, or where access to the developer may be limited.

[2] Whole this section (5.4) is from : Common criteria, ISO/IEC 15408, part: 3; ver3.1:2009.

5.4.4 Evaluation Assurance Level 3 (EAL3) - Methodically Tested and Checked

EAL3 permits a conscientious developer to gain maximum assurance from positive security engineering at the design stage without substantial alteration of existing sound development practices. EAL3 is applicable in those circumstances where developers or users require a moderate level of independently assured security, and require a thorough investigation of the TOE and its development without substantial re-engineering.

5.4.5 Evaluation Assurance Level 4 (EAL4) - Methodically Designed, Tested, and Reviewed

EAL4 permits a developer to gain maximum assurance from positive security engineering based on good commercial development practices which, though rigorous, do not require substantial specialist knowledge, skills, and other resources. EAL4 is the highest level at which it is likely to be economically feasible to retrofit to an existing product line.EAL4 is therefore applicable in those circumstances where developers or users require a moderate to high level of independently assured security in conventional commodity TOEs and are prepared to incur additional security-specific engineering costs.

5.4.6 Evaluation Assurance Level 5 (EAL5) - Semi Formally Designed and Tested

EAL5 permits a developer to gain maximum assurance from security engineering based upon rigorous commercial development practices supported by moderate application of specialist security engineering techniques. Such a TOE will probably be designed and developed with the intent of achieving EAL5 assurance. It is likely that the additional costs attributable to the EAL5 requirements, relative to rigorous development without the application of specialized techniques, will not be large. EAL5 is therefore applicable in those circumstances where developers or users require a high level of independently assured security in a planned development and require a rigorous development approach without incurring unreasonable costs attributable to specialist security engineering techniques.

5.4.7 Evaluation Assurance Level 6 (EAL6) - Semi Formally Verified Design and Tested

EAL6 permits developers to gain high assurance from application of security engineering techniques to a rigorous development environment in order to produce a premium TOE for protecting high value assets against significant risks. EAL6 is therefore applicable to the development of security TOEs for application in high risk situations where the value of the protected assets justifies the additional costs.

5.4.8 Evaluation Assurance Level 7 (EAL7) - Formally Verified Design and Tested

EAL7 is applicable to the development of security TOEs for application in extremely high risk situations and/or where the high value of the assets justifies the higher costs. Practical application of EAL7 is currently limited to TOEs with tightly focused security functionality that is amenable to extensive formal analysis.

ISO/IEC 15408 part: 3 define some assurance requirement. This requirement including evaluation levels which say scale of evaluation, and including Composed Assurance Packages (CAP) that defines scale for evaluation product assurance combination and is a set of assurance components have been chosen as CAP. The cap is a metric for evaluating ST and PP.

6 Security Engineering

Now, we want classified some important concepts of software engineering that help us in implementing secure information system. The category that is showed below is a rotary motion which is begins at early step of information systems development, and raise with information system's life time. In every cyclic motion we can see more security in information systems [3], [6], [9].

6.1 Basic Properties and Requirement to Obtained Security

Completeness: it means to be complete engineering process demands and security requirement engineering and have to be identified correctly and identified all the user needs or employer or the product of system evaluator.

Granularity: which means requirements based on cost and time and investor preferences and functional or non-functional properties of production such as quality (be care that the quality may be deferent for the region or in deferent applications), performance and … would be classified and grouped. Identifying the importance and unimportant requirement group and other systems located in this category.

Integration: against the granularity which separate requirements, it combine requirement and solutions. Attribute with the same solutions combined in this method to allow the instrument to create an engineering process.

Resiliency: it means information system (IS) communicated with other systems in a secure platform with secure communicate infrastructure.

Quality: Quality defines in two topics: Functional and non-functional. In the first case the special software engineering test is produced and in the second case definition has been obtained depending on organizational needs and defining specific metric for quality.

Integrity: it is unity of security policy that will meet requirement's group and cause to resist in any changes and unauthorized penetration from unauthorized users.

Confidentiality: all solution to forbid unauthorized users and unauthorized access an unauthorized query even authorized user by some concepts such as encryption, encapsulate or stenography … .

Availability: all solution that permit to authorized user to access authorized source in authorized time and authorized place.

6.2 Lateral and Complementary Requirement and Properties for Security

Non-Repudiation: all of the operations have to recorded and stored, and all factors connected to the system have to be documented and justified.

Accountability: it means each operation of any operator on system has to be able tracing it.

Dependability: this property is definitely been established if two basic properties 'availability and integration' implemented before.

Predictability: we have to considered all aspects and all possible state to penetration and attacks, if the inability to resolve deficiency and vulnerabilities we have to predict damages and potential risks.

Basically the other main and complementary requirement noted above, not to be as an absolute process. There is a cyclic process for implementing security. That are always started of completeness, granularity and …, and is obtain than the level of security at each iteration, this cycle repeat and each repeat will implement and evaluate higher and better level of security and impractical individuals little by little. Above concepts are illustrated on figure 4:

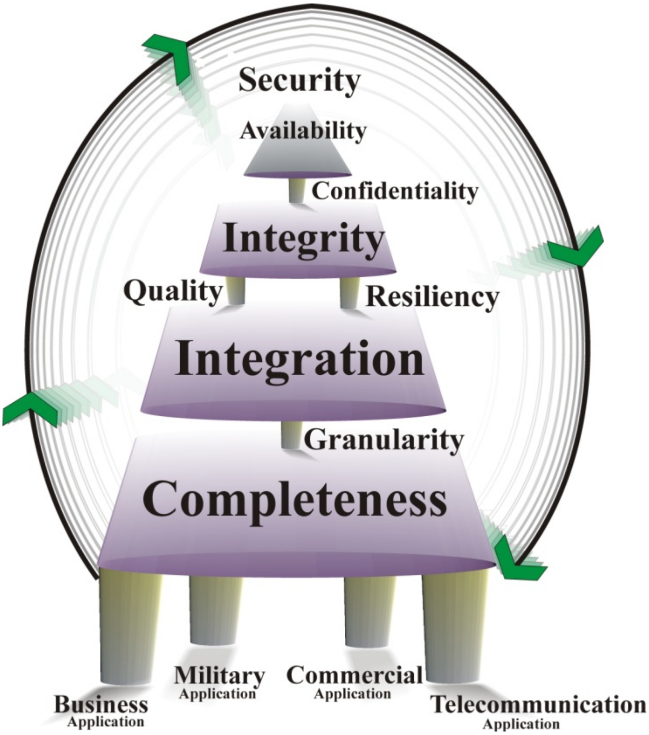


Fig. 4. Basic properties and requirement to obtained security. With a rotary motion that use security evaluation and help us to implement secure IS.

According to all stages of product life cycle (PLC) and software life cycle (SLC) suggested in figure 2, improved from security designing and development (secure production process) point of view and considered same discussion such as secure design, secure coding … as an engineering process. As we see in security requirements engineering, there are some methodologies such as SQUARE and etc.

With proper understanding of process suggested in figure 2 and 3, we will see the complex concept of security will change to a repeatable, well defined, refined ability, optimization process and achieved away from the complexities of a much lower cost and greater reliability and reliance, step by step and obviously to implementation, maintenance, review and improving information systems security.

Result

Evaluation has two types: Evaluation of final product and evaluation of process of producing. The CC uses both of them with definition of some security component and security evaluation levels, so it has become a comprehensive and acceptable standard for all the organizations and institutions which work in field of security.

The CC has encouraged user to emphasis on importance of whatever has been said below:

- Identify requirement of secure information systems and require to security policy and secure goal setting for secure information systems.
- Implementation and exploitation of controlled cases in order to manage secure information system risks.
- Reviewing of performance and effectiveness of secure information system.
- Continuous improvement based on measurement and evaluation of secure goals.
- Harmony with other management systems standards.

Our proposal is based on a process which help to establish, perform, exploit, refined secure property, review the information management systems and show the way to maintain, improvement and development this process. The implementing of such a system has to be as a strategic decision for an organization's directors. Among these there are effective factors on this strategic decision such as: security requirement, organization's security goals, process applied and size and structure of organization.

The directors should consider the security as a result of rotary engineering process and they should understand the gradual process for achieving security. In this rotary motion, a cyclic reproducibility engineering mechanism, will cause to remove barriers then may make the optimal conditions to produce secure Information system (IS) in terms of time and energy an investment and cause to more secure products.

According to security policy and security plans of organization which have made based on ISO/IEC 15408 and have determined as an organization guideline, directors must have different teams in different phases of Information Systems life cycle who those aren't particularly security specialists and these teams would determine steps with documentary of their attempts and experiences and compete with secure aims, so they can perform to product implementation in secure form from the beginning of process instead of performing security at the end of producing, and it will have less cost and more approaches. And they will success to produce secure product and they will be able to create the secure produce of product process step by step.

Acknowledgment

Some of the materials collected from different websites.

References

1. Common criteria, ISO/IEC 15408, part: 1 - 2- 3; ver3.1 (2009)
2. Carnegie Mellon University's Software Engineering Institute, http://www.cert.org
3. ISO 27000 standards, http://www.27000.org
4. ISO 15408, http://www.groups.27000.ir/iso15408
5. Firesmith, D.G.: Engineering Safety & Security-Related Requirements for SW-Intensive Systems. In: 32nd International Conference on Software Engineering. Carnegie Mellon University, Pittsburgh (2010)
6. Mellado, D., Fernandez-Medina, E., Piattini, M.: Security requirements engineering framework for software product lines. Information and Software Technology 52, 1094–1117 (2010)
7. Mead, N.R., Allen, J.H., Ardis, M., Hilburn, T.B., Kornecki, A.J., Linger, R., McDonald, J.: Software Assurance Curriculum Project Volume II: Undergraduate Course Outlines. Technical Report Cmu/Sei–TR-005, ESC-TR–005 (2010)
8. Mead, N.R., Allen, J.H., Arthur Conklin, W., Drommi, A., Harrison, J., Ingalsbe, J., Rainey, J., Shoemaker, D.: Making the Business Case for Software Assurance. 78, Special Report Cmu/Sei-2009-SR-001 (2009)
9. Mellado, D., Fernandez-Medina, E., Piattini, M.: Towards security requirements management for software product lines: A security domain requirements engineering process. Computer Standards & Interfaces 30, 361–373 (2008)

Computer Security Threats Towards the E-Learning System Assets

Zainal Fikri Zamzuri, Mazani Manaf, Adnan Ahmad, and Yuzaimi Yunus

Faculty of Computer and Matematical Science,
Universiti Teknologi MARA (UiTM),
Shah Alam, Selangor, 40450, Malaysia
zfikri@melaka.uitm.edu.my

Abstract. E-learning system is a web-based system which is exposed to computer threats. Services or asset of the e-learning system must be protected from any computer threats to ensure the users have peace of mind when using it. It is important to identify and understand the threats to the system in order develop a secure system. The main objectives of this paper are to discuss the computer security threats towards the e-learning system assets and to study the six categories of computer security threats to the e-learning assets. The activities which involve the e-learning assets will be analyzed and evaluated using the STRIDE model. The results show that the e-learning system assets are exposed to threats on availability, integrity and confidentiality .Our findings also show that the high risk assets are assessment and students' assessment marks.

Keywords: E-Learning, Computer Security, E-Learning Security, STRIDE, Data Classification.

1 Introduction

Nowadays, e-learning system has becomes popular among the educational institutions. This is because E-learning system gives a lot of benefits to people such as guaranteed 24-hour response to student questions, education taking place anytime, anywhere and searchable knowledge base. E-Learning is also quite effective in its ability to produce measurable results by monitoring attendance, effectiveness, performance, and recording test scores [1].

Since the e-learning system is run under internet environment, therefore it is exposed to computer threats and vulnerabilities of internet. Threat is an impending action by a person or event that poses some danger to assets. A loss of an asset is caused by the realization of threat. The threat is realized via the medium of vulnerability [2]. It is very important to know and understand all the threats towards the system to be developed. In risk management process, the evaluation of risk on assets, threats and vulnerabilities are done in assessment phase [3]. Security risk analysis, otherwise known as risk assessment, is fundamental to the security of any organization. It is essential in ensuring that controls and expenditures are fully commensurate with the risks to which the organization is exposed. The objective of risk analysis is to identify

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 335–345, 2011.

risks from potential events with a view to reduce the level of risk to an acceptable level [4].

Assets are valuable resources of the organization that need to be protected. The loss of assets represents the significant loss to the organization. In some cases, a lost asset cannot be replaced particularly in the case of goodwill, trust, or confidential research. Examples of asset categories are; users, services, servers, networks, documentation, goodwill, reputation and personnel skills [2].

STRIDE is one of the methods to identify all the possible threats towards the system by classifying the threats into six categories which are Spoofing, Tampering, Repudiation, Information disclosure, denial of service and Elevation of privilege. Swiderski and Snyder (2004) suggested that threats can be classified into six classes based on their effect [5].

2 Related Research

There are a few research had been done about the e-learning security. None of the research had discussed computer security threats towards the e-learning system assets as discussed in this paper. Most of the researchers discussed a security issues on e-learning system related to specific area such as privacy, protecting the e-learning content, authentication and on-line assessment.

Privacy in the e-learning system had been discussed by [6], [7]. [8], [9] had discussed on how to protect the e-learning content from being used without permission. [10], [11], [12] had discussed the authentication system for the e-learning system and on-line assessment had been discussed by [13], [14], [15], [16], [17]. Maria et al. [18] had discussed different types of availability, integrity and confidentiality attack on the e-learning system. Yong (2007) discussed the security attributes that are relevant to all e-learning stakeholders and suggested a security modeling for e-learning system that describes the relationships among e-learning stakeholders [19].

3 E-Learning Assets

Shareable Content Object Relational Management (SCORM) defines an asset as a simple resource, such as a static HTML page or a PDF document, or collection of files, such as images and a style-sheet [6]. Whereas [7] have looked at asset of e-learning system in a more specific way , where they defined e-learning assets as E-Learning content (Exam, Notes, Grade),Cryptographic key content, User personal data, Messages between users, Different group membership data, Network bandwidth, Message integrity and Message availability.

In this discussion, writers will define e-learning asset as services provided by e-learning system such as learning resources, examination or assessment questions, students' results, user profile, forum contents, students' assignment and announcement in the e-learning system.

3.1 Learning Resources

Learning resources are assets that provide students with lecture notes to help students in their studies. Learning resources are uploaded by the facilitator for their students. Students hope that the learning resources such as lecturer notes that they download from the system are not changed from the original content. The facilitators also hopes that their notes distributed to students are not changed by unauthorized people and distributed to others without their knowledge. The facilitators want their copyright on their lecture notes. Students will not feel happy if the learning resources uploaded into the system is unavailable for downloading when they need them.

Weippl (2005) stressed that all information provided by the university's e-learning system must be free of errors to avoid damage to the reputation and prestige of individual departments and the university [8].

3.2 On-Line Assessment

The privacy, integrity and availability of these assets have to be really guarded carefully. Weippl (2005) stressed that security requirements such as integrity, availability, confidentiality and non-repudition of assessment are major factors that influence the success of the on-line assessment [8]. The exam questions and student answer sheet have to be protected from tampering to ensure the integrity of the examination. Students should not know the question before the exam is conducted to ensure the confidentiality of the examination. The system has to be protected from any action to crash the system when the examination is running to ensure the availability of the examination. Non-repudiation is important as evidence of students taking the examination and submitting the answer sheet. The system must only allow registered students to sit for the assessment.

The system should also be able to detect any cheating action during examination conducted such as copying.

3.3 Students' Results

This asset keeps the information about student's performance such as continuous assessment, assignment and examination result. This information should be known by the owner only. This asset can only be accessed by the student and the facilitator. The facilitator will key in and update the information of this asset.

Unauthorized modification of this information will result in the loss of integrity and if someone else knows this information it will result in loss of privacy.Wrong keying-in of students' marks will also affect the integrity of students' marks.

3.4 User Profile

Profiles of students, facilitators and administrators will be keyed-in by the administrator. The student and facilitator can only update certain profiles themselves once their records already exist in the system. Although the information is not as sensitive as examination questions, this information should be protected to safeguard the privacy of users. Accuracy of students' and facilitators' information is also important because any wrong information about them can have adverse effects.

3.5 Forum Contents

This service is used by students and facilitators for the discussions. Students can send questions and await responses from other students or facilitators. Some of the discussion probably involves a sensitive issue; therefore the privacy of this forum must be protected. Forum discussion should enable anonymous postings, because some students would not publish controversial topics if their identity could be revealed [8]. Each message sent to this forum should be tracked through the log files to prevent repudiation among users.

3.6 Announcement

This service is used by the administrator and facilitator to disseminate information to the user especially to the students. The information is not really sensitive information since wrong information does not really affect the organization and there is still space to make amendments.

3.7 Students' Assignment

Students submit their assignment by uploading their work into the system. Student feels happy if their assignment is not modified or tampered when their facilitator receives it. The e-learning system has to maintain the availability of the system especially when the due date is near. Student will feel frustrated if they cannot submit their assignment because of the unavailability of the system. All the assignments submitted to the facilitator needs to be proven to avoid repudiation.

4 Methodology

In this study, all the activities in e-learning system that involves the assets will be analyzed and evaluated using the STRIDE Model. STRIDE model will identify all threats to the assets and categorize them into six categories, if exist. The threats identified will then be classified to determine the risk of threats to those assets. This process will use data classification as discussed below.

4. 1 Data Classification

All assets of e-learning systems will be classified into three categories to determine the security controls to protect these assets. [9], [10], [11] have suggested restricted, confidential and public as three categories of data classifying as shown in table 1.

Table 1. Data Classifying

Category	Description	Level
Restricted	Restricted data will cause significant impact to the organization when the data been unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access. Only those individuals with explicit authorization are designated for approved access. Examples of restricted data are credit card number and bank accounts.	3
Confidential	Confidential data will cause negative impact to the organization when the data has unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access. Other employees and non-employees who have business need to know, delegate access privileges. Examples of confidential data are salary and project quotation.	2
Public	Public data will cause little or no risk impact to the organization when the data has unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access. Organization affiliates public with a need to know. Examples of public data are company profile and company organization chart.	1

4.2 STRIDE

STRIDE is a classification scheme for characterizing known threats according to the kinds of exploits that is used (or motivation of the attacker) [12]. The STRIDE acronym is formed from the first letter of each of the following categories [13].

1. Spoofing: Whenever the communication line between the web service consumer and provider crosses a trust boundary, there is a threat of spoofing.
2. Tampering: Tampering can be done while data is on the communication channel, while data resides on the consumer machine, or while it resides on the provider machine.
3. Repudiation Users may dispute transactions if there is insufficient auditing or record keeping of their activity
4. Information disclosure: Information can leak during communication, or while being stored on consumer or provider machine
5. *Denial of service*: Denial-of-service attacks try to disturb the services by overloading the communication line, or by enforcing a crash or ungraceful degradation of the consumer or provider.
6. *Elevation of privilege*: An elevation of privilege can occur on both the consumer's and producer's machine.

Each of the STRIDE categories can be related to the security objectives. The spoofing and elevation of privilege threats will affect the system authorization, tampering threat will affect the asset integrity , information disclosure threat will affect the confidentiality of the asset and denial of service threat will affect the availability of the asset.

5 Finding and Analysis

All e-learning assets are exposed to security threats. Each of the threat gives a different risk impact to e-learning asset as shown below.

5.1 Learning Resources

Spoofing

- Unauthorized user will be able to read and download the learning resources.
- Unauthorized user with facilitator id will able to delete and modify existing learning resources and uploading unnecessary learning resources.

Tampering

- Registered user will get a wrong knowledge from the learning resources.
- The image of the lecturer and organization will be tarnished for supplying low quality learning resources.
- Student will not thrust the learning resources supplied by the system.

Repudiation

- The facilitators denied upload new learning resource, modify and remove the existing learning resource.

Information disclosure

- Unauthorized user will get learning resources for free.

Denial of service

- The students cannot access and download the learning resources.
- The facilitator cannot upload and update the learning resource.

Elevation of privilege

- Users who have privilege on this asset will be able to remove or modify the available learning materials. They can also upload the unnecessary materials to the system.

This asset has not really affected the organization and the user if the unauthorized people are able to see the learning material since it is not really confidential. However, if the learning materials are being modified without permission it will affect the students' performance as well as damage of the reputation and prestige of individual department and the entire university if it continuously happens.

Therefore, the sensitivity of this asset can be classified as confidentiality or level 2.

5.2 On-Line Assessment

Spoofing

- The unauthorized user can take the exam on behalf of other student.
- If the intruder uses the lecturer's id, the unauthorized user can modify and remove the assessment question paper.

Tampering

- A tampering examination paper can make the examination not valid. A new examination has to be redone and this can affect the student mentality and tarnish the institution image.
- The submission of question paper by student been modified from the original. This can affect the student performance.

Repudiation

- Student admitted that they took the examination even they are not.
- Students who do not submit the examination paper, admits that he/she has already submitted the exam paper.
- Lecturer denies uploading the examination paper.
- Lecturer denies receiving the student answer paper.
- Lecturer denies modifying or removing the existing examination question paper.

Information disclosure

- The examination question is already been known before the examination. The examination is not valid.
- The answer of the examination is known before the examination is completed.

Denial of service

- The examination cannot be held because the system is not available.
- Student cannot submit the exam paper because the system is not available.
- Lecturer cannot upload the question paper for the examination.

Elevation of privilege

- User who has the privilege on this asset will be able to know, modify, add and remove the question paper and student's answer sheet.

The unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will cause significant impact to the organization and users. The University has to rerun the examination if the examination is not valid because the examination question paper or student's answer sheet has been tampered or the answer is already disclosure to the student. This not only will involve extra cost to the university and student but also will damage the image of the university. Therefore, the sensitivity of this asset can be classified as restricted or level 3since it will cause significant impact to the organization when the data is unauthorized for disclosure, alteration or destruction by unauthorized people and disruptions of access.

5.3 Students' Results

Spoofing

- Intruder can upload, remove and edit the marks.
- Intruder can see the marks.

Tampering

- Assessment marks which are modified will make the assessment marks invalid and can affect the student's overall results.

Repudiation

- Lecturer denies uploading and editing the students' assessment marks.

Information disclosure

- Students will lose their privacy

Denial of service

- Students cannot check theirs marks.
- Administrator and facilitator cannot upload students' marks and evaluate the students' performance.

Elevation of privilege

- User who has privilege to this asset will be able to check, add, modify and remove the students' assessment marks.

The unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will give negative impact to the organization and the users. Therefore, the sensitivity of this asset can be classified as confidential or level 2.

5.4 User Profile

Spoofing

- Some confidential information the user will leak to the unauthorized user.

Tampering

- Users' profiles that are modified will give a bad impact to the user.

Repudiation

- Users deny modifying their profile.

Information disclosure

- -

Denial of service

- Users cannot update and add their profiles.

Elevation of privilege

- The user who has privilege to this asset will be able to add, modify and remove the users' profile.

Since the university is interested only in the users' profile, the unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will cause less risk or no impact to the organization and the users. Therefore, the sensitivity of this asset can be classified as public or level 1

4.5 Forum Contents

Spoofing

- Unauthorized user will be able to send message to forum.

Tampering

- The forum content is tampered such as student posting or feedback from facilitator.

Repudiation

- Student or facilitator denied posting questions or feedback on the forum.

Information disclosure

- Since some of the discussion is confidential among the student and facilitator, the disclosure of the forum content can make the student lose confidence in the discussion. This can limit their discussion.

Denial of service

- Discussion among the users cannot be done and user cannot share ideas
  Elevation *of privilege*
- User who has privilege to this asset will be able to check , add, modify and remove the contents in the forum

The unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will cause less risk to the organization and the users. Therefore, the sensitivity of this asset can be classified as public or level 1.

5.6 Students' Assignment

Spoofing

- -

Tampering

- Tempered assignment will affect the student's performance.
- Student will answer the wrong question.

Repudiation

- Student admits he/she has submitted the assignment.

Information disclosure

- -

Denial of service

- Student cannot upload the assignment.
- Facilitator cannot download the student's assignment for marking.

Elevation of privilege

- User who has privilege to this asset will be able to modify, remove and look at into the student assignment.

The unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will have serious impact to the organization and the users. Therefore, the sensitivity of this asset can be classified as confidential or level 2.

5.7 Announcement

Spoofing

- -

Tampering

- User will get wrong information.
- Activity has to be cancelled because of the wrong information.

Repudiation

- Nobody is accountable to the announcement posted.

Information disclosure

- -

Denial of service

- User cannot get the latest information.
- Administrator and facilitator cannot disseminate the latest information to the user.

Elevation of privilege

- User who has privilege to this asset will be able to add, modify and remove the announcement.

The unauthorized disclosure, alteration or destruction by unauthorized people and disruptions of access of this asset will cause less risk to the organization and the users. Therefore, the sensitivity of this asset can be classified as public or level 1.

6 Conclusion

The results show that the e-learning system assets are exposed to threats on availability, integrity and confidentiality. Our finding also shows that the high risk assets are assessment and students' assessment marks. The mitigation risk action has to be taken to protect the high risk assets.

Authentication, encryption and firewall system are the suggested methods to deal with computer security threats towards the e-learning systems assets. The problem is to select the best techniques which are suitable for the system since there are a lot of techniques available.

Further research is needed to rank the threat, identify the attacks that cause the threats and to identify the countermeasures for each threat.

References

1. Tastle, W., White, B., Shackleton, P.: E-Learning in Higher Education: The Challenge, Effort, and Return on Investment. International Journal on E-Learning 4, 241–251 (2005)
2. Singh, B.: Network Security and Management, vol. 1. Prentice-Hall of India Pvt. Ltd., New Delhi (2007)
3. Gehling, B., Stankard, D.: Ecommerce Security. In: Information Security Curriculum Development (Infoseccd) Conference Kennesaw, GA,USA, pp. 32–37 (2005)
4. Peltier, T.: Information Security Risk Analysis. CRC Press, Boca Raton (2005)
5. Myagmar, S., Lee, A.J., Yurcik, W.: Threat Modeling as a Basis for Security Requirements. In: Proceedings of the Symposium on Requirements Engineering for Information Security (SREIS 2005), Paris (2005)

6. Weippl, E., Tjoa, A.: Privacy in E-Learning: Anonymity, Pseudonyms and Authenticated Usage. International Journal of Interactive Technology and Smart Education (ITSE) 2, 247–256 (2005)
7. Klobu Ar, T., Jenabi, M., Kaibel, A., Karapidis, A.: Security and Privacy Issues in Technology-Enhanced Learning. In: Cunningham, P., Cunningham, M. (eds.) Expanding the Knowledge Economy: Issues, Applications, Case Studies. IOS Press, Amsterdam (2007)
8. Sun, L., Wang, H., Li, Y.: Protecting Disseminative Information in E-Learning. In: Advances in Web Based Learning, ICWL 2007, pp. 554–565 (2008)
9. Graf, F.: Providing Security for elearning. Computers and Graphics (Pergamon) 26, 355–365 (2002)
10. Asha, S., Chellappan, C.: Authentication of E-Learners Using Multimodal Biometric Technology. In: International Symposium on Biometrics and Technology, Islamabad (2008)
11. Agulla, E., Castro, L., Mateo, J.: Is My Student at the Other Side? Applying Biometric Web Authentication to E-Learning Environments. In: Proceedings of Eighth IEEE International Conference on Advanced Learning Technologies (ICALT 2008), pp. 551–553, (2008)
12. Inaba, R., Watanabe, E., Kodate, K.: Security Applications of Optical Face Recognition System: Access Control in E-Learning. Optical Review 10, 255–261 (2003)
13. Marais, E., Argles, D., Von Solms, B.: Security Issues Specific to E-Assessments. The International Journal for Infonomics Special issue: 'e-Learning Security' (2006)
14. Levy, Y., Ramim, M.: A Theoretical Approach for Biometrics Authentication of E-Exams. In: The 2007 Chais Conference on Instructional Technologies Research (2007)
15. Apampa, K., Wills, G., Argles, D., Marais, E.: Electronic Integrity Issues in E-Assessment Security. In: Eighth IEEE International Conference on Advanced Learning Technologies, pp. 394–395 (2008)
16. Weippl, E.: On the Use of Test Centers in E-Assessment, E-Learning Reports. Vienna University of Technology (2006)
17. Hernández, J., Ortiz, A., Andaverde, J., Burlak, G.: Biometrics in Online Assessments: A Study Case in High School Students. In: Proceedings of the 18th International Conference on Electronics, Communications and Computers (Conielecomp 2008), pp. 111–116 (2008)
18. Nickolova, M., Nickolov, E.: Threat Model for User Security in E-Learning Systems. International Journal Of Information Technologies and Knowledge, 341–347 (2007)
19. Yong, J.: Security Modelling for E-Learning. In: Proceedings of the 2007 1st International Symposium on Information Technologies & Applications in Education (ISITAE 2007), Kunming, pp. 1–5 (2007)
20. Ostyn, C.: In the Eye of the SCORM: An Introduction to SCORM 2004 for Content Developers (2007) (retrieved)
21. Weippl, E.: Security in E-Learning (Advances in Information Security). Springer, New York (2005)
22. Olzak, T.: A Practical Approach to Threat Modeling, `http://adventuresinsecurity.com/blog/wp-content/uploads/2006/03/A_Practical_Approach_to_Threat_Modeling.pdf`
23. The UM-Miller School of Medicine's Department of Information Technology, `http://it.med.miami.edu/x1297.xml`
24. M.U. Information Services and Technology, http://ist.mit.edu/security/data_classification
25. Prasath, V.: Modeling the Evaluation Criteria for Security Patterns in Web Service Discovery. International Journal of Computer Applications IJCA 1, 53–60 (2010)
26. Desmet, L., Jacobs, B., Piessens, F., Joosen, W.: Threat Modeling for Web Services Based Web Applications. DistriNet Research Group, Katholieke Universiteit Leuven (2005)

User Acceptance for Extended Function Point Analysis in Software Security Costing

Nur Atiqah Sia Abdullah[1], Rusli Abdullah[2],
Mohd Hasan Selamat[2], and Azmi Jaafar[2]

[1] Faculty of Computer and Mathematical Sciences, University Technology MARA,
43300 Shah Alam, Selangor, Malaysia
[2] Faculty of Computer Science and Information Technology, University Putra Malaysia,
43000 Serdang, Selangor, Malaysia
atiqah@tmsk.uitm.edu.my, {rusli,hasan,azmi}@fsktm.upm.edu.my

Abstract. This paper explains the user acceptance used in evaluating the Extended Function Point Analysis (Extended FPA) in software security costing. The construct of Software Security Characteristics Model (SSCM), the validation of SSCM, prototype as well as adaptation of the user acceptance models, are discussed in this paper. This paper also emphasize on the user acceptance test for the prototype. The experiment construct includes the variables selection, subject selection, hypotheses formulation, and treatment. User acceptance questionnaire is setup followed by the experiment. Results show that Extended FPA is perceived ease to use, more useful as well as more likely to use, rather than IFPUG FPA in calculating software security cost.

Keywords: Software security cost, function point analysis, Software Security Characteristics Model (SSCM), user acceptance, Technology Acceptance Model (TAM), Method Evaluation Model (MEM).

1 Introduction

Function Point Analysis (FPA) is an ISO recognized Functional Size Measurement (FSM) method. It is currently maintained by International Function Point User Group (IFPUG). Therefore, it is commonly known as IFPUG FPA. FPA is invented by Albrecht [1] to overcome the sizing problem caused by Source Line of Code (SLOC) method. It is also one of the widely used software cost estimation (SCE) method.

Albrecht's FPA [1] suffers from some essential problems and passes many stages of evolution to solve these problems [2]. Many researchers created various releases of function point (FP) measure such as Feature Points [3], Mark II FPA [4], 3D FP [5], Full Function Point (FFP) [6], COSMIC FFP [7] and etc.

Costs related to computer security are often difficult to assess in part because accurate metrics have been inherently unrealistic [8]. Finding and eradicating software defects early is cost-effective and economically sane [9]. For example, fixing a defect post-production takes on average 32 hours of development time [9]. Costs that more difficult to quantify but have resulted in severe loss of use or productivity

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 346–360, 2011.

include viruses and malware, web server denial-of-service attacks, abuse of access privileges and equipment vandalism [10]. Integrating controls and best practices into an existing software development life cycle (SDLC) is needed [9].

The security issues in project attributes are not considered and unable to determine in most of the models, include SLIM, checkpoint, Price-S, Estimacs, SELECT Estimator and COCOMO II [11]. Neither the COCOMO II nor the COCOTS estimating model includes security as a cost driver [12]. Many commercial firms are interested in determining how the costs of implementing different security strategies [13]. Therefore, COCOMO II is extended to COSECMO [14], Security Cost Driver [12], and Security Risk Analysis [15]. All these research focus on COCOMO II. In previous works, the most commonly used parametric FSM methods such as IFPUG FPA, Mk II FPA, COSMIC FFP, SLIM, and COCOMO II, are evaluated from the security viewpoints [16].

In our study, FPA is selected to be extended to estimate the security costing. It is due to the widely usage of FPA in the SCE. We also proposed Software Security Characteristics Model (SSCM) to be extended in the General System Characteristics (GSCs) calculation [17]. However, the user acceptance is not yet being carried out for the proposed model. Therefore, in this paper, a laboratory experiment is carried out to evaluate the user acceptance for the proposed SSCM, which has been extended in the FPA. This experiment results compared the performance behavior of the Extended FPA with IFPUG FPA.

2 Related Works

This section elaborates the related works on the Software Security Characteristics Model (SSCM), model validation using Rasch measurement analysis, prototype, adaptation of user acceptance models, and lastly the validation of prototype through user acceptance test.

2.1 Software Security Characteristics Model

Software Security Characteristics Model (SSCM) [16][18] is an integration of two software security metrics with four common security standards. The software security metrics are Software Security Management and Metrics [19] and McGraw's Software Security Seven Touch Points [20]. In this integration, there are four common security standards; namely, Information Technology Security Cost Estimation Guide [21], Common Criteria for Information Technology Security Evaluation [22], Open Web Application Security Project [23], and Control Objectives for Information and related Technology [24].

SSCM has proposed five basic phases in the SDLC, which consists of Plan (P), Design (D), Code (C), Test (T), and Deploy (E). In each phase, there are interrelated security aspects. These security aspects are selected from two software metrics. The security aspects are arranged in SDLC as shown in Table 1.

Table 1. Proposed Software Security Aspects

Step	Security Aspects
Plan (P)	Security Requirements (SR)
Design (D)	Security Features (SF); Functional Features (FF)
Code (C)	Attack Planning (AP); Formal Review and Sign-off (FR); Secure Coding, Review and Audit (SCR)
Test (T)	Software Security Assurance (SSA); Final Security Review (FSR); Infrastructure Application Security Measures (ASM)
Deploy (E)	Software Hardening & Application Security Monitoring (SHA)

The security aspects are cross-referenced with four common security standards as in Table 2.

Table 2. Cross-reference of Four Common Security Standards

Security Aspects in SSCM	ITSCE	CCITSE	OWASP	COBIT
Security Requirements (SR)	√	√	√	√
Security Features (SF)	√	√	√	√
Functional Features (FF)	√	√	√	
Attack Planning (AP)			√	
Formal Review and Sign-off (FR)	√	√		
Secure Coding, Review and Audit (SCR)	√		√	√
Software Security Assurance (SSA)		√		
Final Security Review (FSR)				√
Infrastructure Application Security Measures (ASM)	√			√
Software Hardening & Application Security Monitoring (SHA)	√	√	√	√

There are 48 software security characteristics, which are derived from these security aspects [25].

2.2 Validation of SSCM

SSCM is validated through a survey with Malaysian Multimedia Super Corridor (MSC) software developers. The collected data are analyzed using Rasch measurement method. Rasch measures the competency in an appropriate way to ensure valid quality information can be generated for meaningful use; by absorbing the error and representing a more accurate prediction based on a probabilistic model [26].

From the analysis, generally the respondents have high level of awareness in implementing the software security characteristics in SSCM throughout SDLC; μ_{person} of 83.06%, which is higher than 60% threshold limit. Hereby, the Person Mean = $1.59 \geq 0.00$; with significant of $p=0.05$. Therefore the H_0 is accepted. From this survey, the level of awareness of software security characteristics throughout SDLC in SCE among software developers is 83.06%, where suggested that these security characteristics are valid, relevant and implemented in current practices [27].

2.3 Prototype

From the analysis of the survey, some enhancements are made in the GSCs calculation of IFPUG FPA [17]. There are 14 GSCs in IFPUG FPA. The summation of GSCs is called Value Adjusted Factor (VAF). The enhanced formula is as follows:

$$\text{VAF} = 0.65 + [(\sum_{i=1}^{14} Ci + \text{Security}) /100]. \qquad (1)$$

where

Ci = degree of influence for each GSC
I = is from 1 to 14 representing each GSC
Σ = is summation of all 14 GSCs
Security = Degree of Influence for Security Characteristics.

The degree of influence (DI) for security characteristics is evaluated through two additional evaluation sheets [17]. These evaluation sheets are produced based on the SSCM to help the user to estimate the security costing. A prototype is designed based on this study. It is named as Extended Function Point Analysis Prototype [28]. In this paper, this prototype is referred as extended FPA.

3 User Acceptance Models

To investigate the user acceptance on this extended FPA, user acceptance models are used to predict the likelihood to adopt in practice. There are two acceptance models highlighted in this paper. These models are then adapted with ISO/IEC to identify the dependent variables in a laboratory experiment.

3.1 Technology Acceptance Model

The Technology Acceptance Model, TAM [29] is one of the most frequently tested models in Management Information System (MIS) literature. TAM attempts to predict and explain computer-usage behavior. TAM was derived from the Theory of Reasoned Action (TRA) [30], which a person’s performance of a specified behavior is determined by his/her behavioral intention to perform the behavior; and the behavioral intention is jointly determined by the person’s attitude and subjective norms concerning the behavior in question [30]. TAM uses perceived ease of use and perceived usefulness of the technology as two main determinants of the attitudes toward a new technology [29].

3.2 Method Evaluation Model

The Method Evaluation Model (MEM) [31] is a theoretical model for evaluating information system (IS) design methods, which incorporate both aspects of method in success: actual performance and likely adoption in practice. MEM combines Rescher's Theory of Pragmatic Justification [32] and Davis's TAM [29].

In this context, the core of the MEM, as shown in Figure 1, consists of the same perception-based constructs as the Davis's TAM [29], but now adapted for evaluating methods. These constructs include Perceived Ease of Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU). PEOU shows the degree to which a person believes that using a particular method would be free of effort. PU indicates the degree to which a person believes that a particular method will be effective in achieving its intended objectives. ITU gives the extent to which a person intends to use a particular method. These central constructs are so called the Method Adoption Model (MAM) [31][33].

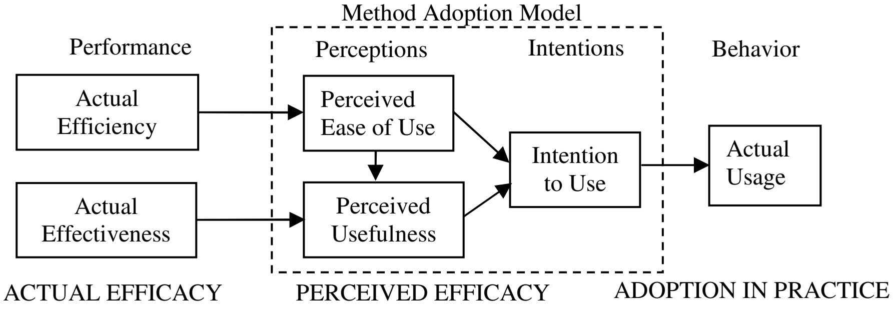


Fig. 1. Method Evaluation Model [31]

This model is extended with additional constructs that provide inputs to the MAM and predict its ultimate output whether the method will be used in practice. The inputs for MAM are Actual Efficiency, which indicates the effort required applying a method; and Actual Effectiveness represents the degree to which a method achieves its objectives. The output variable from MAM is Actual Usage, which shows the extent to which a method is used in practice.

3.3 ISO/IEC 14143-3

In this paper, Part 3 of the Information technology - Software measurement - Functional size measurement in ISO/IEC is included. This part is selected because it contains the verification of the FSM methods [34]. It is used to evaluate the actual performance of a FSM method in this study. The efficiency of a FSM method is defined by the effort required understanding and applying the FSM method. It can be measured using the following measures such as time, cost, productivity and cognitive effort. The effectiveness of a FSM method is defined by how well it achieves its objectives. Effectiveness can be measured using the specific properties and requirements of FSM. These performance properties include repeatability, reproducibility, accuracy, convertibility, discrimination threshold, and applicability to functional domains. Hence, two performance-based variables in the MAM, which are efficiency and effectiveness, are measured. In this experiment, time, reproducibility and accuracy are chosen to evaluate the efficiency and effectiveness of the FSM method.

3.4 Adaptation of User Acceptance Models

TAM, MEM and ISO/IEC are considered and adapted as the user acceptance model in this study. There are two main types of dependent variables based on these models, which are performance-based variables and perception-based variables. For the performance-based variables, variables include time; productivity; reproducibility; and accuracy as shown in Table 3.

Table 3. Performance-based Variables

Variable	Description
Time	Time taken by a subject to complete the sizing task
Productivity	FP count produced by a subject, called FP Size
Reproducibility	Agreement between the measurement results of different subjects using the same method; Difference between assessment values of the subjects.
Accuracy	Agreement between the measurement results and the true value; Difference between assessment values of each subject to true value.

For perception-based variables, three variables are perceived ease of use; perceived usefulness; and intention to use as shown in Table 4.

Table 4. Perception-based Variables

Variable	Description
Perceived Ease of Use	Degree to which a subject believes that using a particular method would be free of effort or less time
Perceived Usefulness	Degree to which a subject believes that a particular method will be effective in achieving its intended objectives or sizing task
Intention to Use	Degree to which an individual intends to use a particular method as a result of his/her perception of the method's performance

The relationship between these dependent variables and user acceptance models are shown in Table 5.

Table 5. Relationship between dependent variables and user acceptance variables

Dependent Variables	Efficiency	Effectiveness	Adoption
Performance-based	Time	Reproducibility; Accuracy	
Perception-based	Perceived Ease of Use	Perceived Usefulness	Intention to Use

Questions in the User Acceptance Questionnaire. In order to counter check the responses of respondents in the experiment, these suggested questions are designed in pair in a post-task survey with 14 close-ended questions, which representing the dependent variables. These items were formulated using a 5-point Likert scale, and opposing statements format. The order of the items was randomized to avoid monotonous responses. From the adapted user acceptance models, the suggested statements for the user acceptance questionnaire are as shown in Table 6:

Table 6. Questions in User Acceptance Questionnaire [31]

Variable	Question
PEOU1.	I found the procedure for the FSM method simple and easy to follow
PEOU2.	Overall, I found the FSM method easy to use
PEOU3.	I found the measurement rules of the FSM method clear and easy to understand
PEOU4.	I found the FSM method easy to learn
PEOU5.	I found it easy to apply the FSM method to the case study
PU1.	I believe that this FSM method would reduce the time required to measure secure systems.
PU2.	Overall, I found the FSM method to be useful
PU3.	I think that this FSM method would improve the accuracy of estimates of secure systems
PU4.	Overall, I think this FSM method does provide an effective way of measuring the functional size of secure systems during the requirements phase.
PU5.	Using this FSM method would improve my performance in measuring secure system.
PU6.	Overall, I think this FSM method is an improvement to the IFPUG FPA method.
ITU1.	I will use this FSM method if I have to measure secure systems in the future.
ITU2.	It would be easy for me to become skilful in using this FSM method.
ITU3.	I intend to use this FSM method in the future.

4 Experiment Construct

A laboratory experiment is carried out to test the user acceptance towards the extended FPA compared to IFPUG FPA in estimating the software security cost. This experiment was guided by the framework for experimentation software engineering [35] and adapted experimentation procedures [36]. The main goal for this experiment was to determine whether Extended FPA or IFPUG FPA is a better functional size assessment method for security costing when measuring the same user requirements. It is also to assess which method has the better performance and likely to be adopt in practice.

To compare whether the Extended FPA is more efficient and/or effective than IFPUG FPA as well as more likely to be adopted in practice, we visualized the following hypotheses based on the research questions for this experiment, as in Table 7.

For the efficiency of both methods, separate timings for both methods are taken. It is necessary because the compared methods in this experiment require different identification and measurement steps based on the SRS to the functional size value. For each subject, time, in work-hours spent to complete the sizing tasks associated with each FSM method in each treatment, is collected.

Table 7. Relationship between dependent variables and hypothesis

Variables	Hypothesis	Description
Time	H1	Extended FPA will take more or as much time as IFPUG FPA
Reproducibility	H2	Extended FPA will produce less or equally consistent assessments than IFPUG FPA
Accuracy	H3	Extended FPA will produce less or equally accurate assessments than IFPUG FPA
Perceived Ease of Use	H4	Participants will perceive Extended FPA to be more or equally difficult to use than IFPUG FPA.
Perceived Usefulness	H5	Participants will perceive Extended FPA to be less or equally useful than IFPUG FPA.
Intention to Use	H6	Participants will be less or equally likely to use Extended FPA than IFPUG FPA

For reproducibility (REP), the effectiveness of a FSM method depends on the reliability of the measurements in assessing the same requirements [36]. The closer the measurement results obtained by different raters, the more effective the FSM method is. Therefore, a formula is used to calculate the difference of each subject assessment value with the average assessment value:

$$\text{Subject Reproducibility (REP}_i\text{)} = \frac{\text{AverageAssessment} - \text{SubjectAssessment}}{\text{AverageAssessment}} \qquad (2)$$

For accuracy, even when obtained measurements for the subjects are identical, they might different from the true value for the actual functional size count. Therefore, the accuracy of the method is needed to counter check the effectiveness of the particular method. Thus, we considered the actual functional size that counted by the researcher as the true value in this experiment. This true value is produced by detail counting using IFPUG FPA manual. Then we compared the true value with the measurements that produced by the subjects to get the magnitude of error (MRE) [36].

$$\text{Magnitude of Error (MRE}_i\text{)} = \frac{\text{ResearcherAssessment} - \text{SubjectAssessment}}{\text{ResearcherAssessment}} \qquad (3)$$

As the post-task survey for a FSM method is applied directly after the sizing task of the particular method, it is hypothesized that the perception of performance will be influenced by the experience of applying the particular method according to the MEM. Therefore, the relationship between the independent variables and the perception-based variables is tested indirectly.

4.1 Selection of Subjects

The subjects that participated in this experiment were eight IT personals in Klang Valley, Malaysia. These IT personals were used as the proxies of practitioners because of the following reasons:

a. Accessibility: the possibility of getting practitioners is difficult due to the constraints of time and cost. Therefore, these personals are chosen from Klang Valley, which were willing to participate and the number of subjects is only a small number. These subjects were chosen for convenience.
b. Similarity: these subjects were aged between 28 to 34 years old. They had similar background, which were graduated from degree of computer sciences and have working experiences in software development. They are currently working as senior programmer, project manager, and system analyst. They have stronger relevancy and were representing the actual practitioners and population under study.

4.2 Software Specification Requirements

Software requirement specification (SRS) is a documentation of requirements for a system of a company. This specification is structured according to the directives given by the standard IEEE Recommended Practice for Software Requirements Specification ANSI/IEEE 830 1998 by referenced to the IEEE Std 830-IEEE Guide to Software Requirements Specifications IEEE Standard Board. The aspects in this specification include the purpose of the system, scope, definitions, product perspective, product functions, user characteristics, constraints, assumptions, functional requirements, Entity-Relationship (ER) diagram, user requirements, software security requirements, and external interface requirements.

There are three SRS for this study. Project Management System (PMS) and Student Management System (SMS) are used in the training task. The working example for the IFPUG FPA training session included a requirements specification document of SMS as well as ER diagram. The working example for the extended FPA included a requirement specification of PMS and ER diagram.

The experimental case study was Employee Management System (EMS). It is used during the sizing task. The EMS is the material for the respondent to evaluate and estimate the FP. The FP counts are then entered by using the Extended FPA.

4.3 Experiment Treatments

The treatment in this experiment is corresponding to the two levels of independent variables: the use of extended FPA versus the use of IFPUG FPA to size a SRS with security specification. The within-subjects design [37] is chosen to carry out the experiment to control for differences in human ability.

The within-subjects design is modified to cancel out some possible learning effects. First is the similarity in the treatments, in this case, the relatedness of both FSM methods, the sequence in which the tests was switched. Secondly is the learning effect due to the fact that the same SRS is used for all the subjects. Once the requirement is used, the order of applying the methods might introduce a confounding effect. Therefore, this learning effect is also cancelled by taking into account the sequence of the tests as a factor. The subjects were randomly assigned to two groups, which equal number of 4 in each group and tests presented in a different order as in Table 8:

Table 8. Modified within-subjects design [37]

Group	Treatment		Observation
Group 1 (n = 4)	Tx_a :	IFPUG FPA	Obs_a
	Tx_b :	Extended FPA	Obs_b
Group 2 (n = 4)	Tx_a :	Extended FPA	Obs_a
	Tx_b :	IFPUG FPA	Obs_b

During the observation (Obs_a and Obs_b), the groups are measured with dependent variables. Time is captured by the system and manually recorded. Reproducibility and Accuracy is calculated by using formula that used the value of Productivity, which is the function point size. Perceived Ease of Use, Perceived Usefulness and Intention to Use are gathered through a user acceptance questionnaire.

For the treatments (Tx_a and Tx_b), the respondents are trained to use IFPUG FPA and Extended FPA in the different order. There are three sessions in each treatment. Each treatment has training session, sizing task session and post survey session. There are two trainings in both treatments, which consist of training for IFPUG FPA and training for Extended FPA. Table 9 shows the sequences of the training sessions and corresponded experimental tasks.

Table 9. Training sessions and corresponded experimental tasks

Group 1	Group 2
Training Session in IFPUG FPA	Training Session in Extended FPA
Sizing Task with IFPUG FPA	Sizing Task with Extended FPA
Post-task Survey for IFPUG FPA	Post-task Survey for Extended FPA
Training Session in Extended FPA	Training Session in IFPUG FPA
Sizing Task with Extended FPA	Sizing Task with IFPUG FPA
Post-task Survey for Extended FPA	Post-task Survey for IFPUG FPA

In both treatments, the respondents are taught how to identify the five components in FPA and the GSCs. The respondents are also exposed to the extended FPA to help them in the calculation. During this training task, they were allowed to refer to the training materials.

After the training sessions, the respondents are given a SRS for Employee Management System (EMS) to be used in sizing task, together with the counting procedure for both methods. In the sizing task session, each experimental group used the extended FPA and IFPUG FPA in a different order. The respondents have to calculate the FP without any assistance. However they are still provided with the measurement guidelines that summarizing the measurement rules of the methods. They used the online estimation tool for the counting part. The difference between two methods is the calculation for the software security characteristics costing.

Finally, in the post-task survey, when the subjects had finished the sizing task for a method, they are required to answer the user acceptance questionnaire. They were asked to complete this questionnaire to evaluate the particular method that they had used.

5 Results and Discussion

The descriptive statistics for the IFPUG FPA and Extended FPA methods measurement time and function point size is shown in Table 10.

Table 10. Experiment Results – Time and FP Size

	Time		FP Size	
	IFPUG FPA	Extended FPA	IFPUG FPA	Extended FPA
Mean	2.5875	2.7563	117.480	121.151
SD	0.26152	0.26246	8.2572	8.5153
Min	2.35	2.45	101.8	104.9
Max	3.00	3.00	126.70	130.70

As this distribution was normal, tested by using the Kolmogorov-Smirnov test and measurement time is a continuous variable, the paired t-test is used to check for a difference in mean measurement time between Extended FPA and IFPUG FPA.

To evaluate the significance of the observed difference, we applied the test with a significance level of 5 %, i.e. α = .05. The result of the test as in Table 11 does not allow the rejection of the H1, meaning that we cannot empirically corroborate that Extended FPA will take less time than IFPUG FPA. In fact, for the data collected, the mean measurement time for IFPUG FPA is significantly lower than that for Extended FPA as in Table 10. The reason could be the subjects take into account the software security characteristics of a system. Consequently, the subjects spent more time in applying all the Extended FPA measurement rules.

The differences in reproducibility assessments obtained using both methods were described using the Kolmogorov-Smirnov test to ascertain if the distribution was normal. As a result, the distribution is normal. We decided to use the paired t-test to check for a difference in mean reproducibility between Extended FPA and IFPUG FPA. The result of the test, as shown in Table 11, allows the rejection of the hypothesis H2, meaning that we can empirically corroborate that Extended FPA produces more consistent assessments than IFPUG FPA.

Table 11. Paired t-test for difference in mean measurement time and reproducibility (α = 0.05)

Dependent variable	Time	Reproducibility
Mean	-.16875	.0000103
SD	.21034	.0000007
Std. Error Mean	.07436	.0000003
95% Confidence Interval of the difference	-.34459 (lower) .00709 (upper)	.0000097 (lower) .0000109 (upper)
t	-2.269	40.242
df	7	7
p-value	.058	.000

Next, we tested hypothesis H3 related to accuracy. The value obtained by the researcher is 107.52 FP size for IFPUG FPA and 110.88 for Extended FPA. The MREi is obtained for both methods and the differences in accuracy assessments obtained were described using the Kolmogorov-Smirnov test to ascertain if the distribution was normal. As the distribution was normal, we used paired t-test to check for a difference in mean MREi between Extended FPA and IFPUG FPA. In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %. The result of the test does not allow the rejection of the hypothesis H3 meaning that we cannot empirically corroborate that Extended FPA produces more accurate assessments than IFPUG FPA. The correlation and t cannot be computed because the standard error of the difference is 0. In other words, both methods produce the same accurate assessments. It is due to the reason that the EMS, which provided during the experiment, gave the same criteria of SSCM. It is the flaw of the provided case study in this experiment.

For the comparative analysis of the likelihood of adoption in practice of the FSM methods, we then tested hypotheses H4, H5, and H6, which related to the perceptions of the FSM methods, in terms of perceived ease of use, perceived usefulness and intention to use. Descriptive statistics for the perceptions of the IFPUG FPA and Extended FPA methods are presented in Table 12.

Table 12. Descriptive statistic for perception variables

Dependent variable	Perceived Ease of Use		Perceived Usefulness		Intention to Use	
	IFPUG FPA	Extended FPA	IFPUG FPA	Extended FPA	IFPUG FPA	Extended FPA
Mean	3.425	4.00	3.4813	3.9375	3.3313	3.7500
SD	0.2712	0.321	0.25958	0.27830	0.30861	0.29655
Min	3.0	4.0	3.00	3.67	3.00	4.00
Max	4.0	4.0	3.67	4.33	3.33	4.00

The mean values obtained show that Extended FPA has a higher mean score than IFPUG FPA, meaning that is it perceived to be easier to use, perceived usefulness, and intention to use than IFPUG FPA.

In order to evaluate the significance of the observed difference, we applied a statistical test with a significance level of 5 %, i.e. $\alpha = .05$. As the Kolmogorov-Smirnov test were normal, we decided to use paired t-test to evaluate the statistical significance of the observed difference in mean perceived ease of use, perceived usefulness and intention to use. The result of the test, as shown in Table 13, allows the rejection of the hypothesis H4, H5 and H6, meaning that we empirically corroborate that the participants perceived Extended FPA as easier to use and more useful than IFPUG FPA, as well as the participants more likely to use Extended FPA than IFPUG FPA in calculating software security cost.

Table 13. Paired t-test for difference in mean perception based variables (α = 0.05)

Dependent variable	Perceived Ease of Use	Perceived Usefulness	Intention to Use
Mean	-.5750	-.45625	-0.41875
SD	.4062	0.30430	0.34540
Std. Error Mean	.1436	0.10759	0.12212
95% Confidence Interval of the difference	-.9146 (lower) -.2354 (upper)	-0.71065 (lower) -0.20185 (upper)	-0.70751 (lower) -0.12999 (upper)
t	-4.004	-4.241	-3.429
p-value	.005	0.004	0.011

6 Conclusion

FPA is one of the widely used SCE methods. Evolution from FPA has created various versions of improved FP methods. However, the costs related to computer security are still remained as issues to the estimation. In previous works [16][17][18][25][27], Software Security Characteristics Model (SSCM) is proposed and evaluated. SSCM is extended in the FPA and developed as a tool [28] to estimate the security cost. However, the user acceptance of the estimation tool has to be carried out. This paper focused on the user acceptance models and experiment to evaluate to what extent the user acceptance towards the tool. From the experiment, the responses to the post-task surveys suggest that Extended FPA is more useful and is more likely to be used in the future. The experiment results also show that the participants perceived Extended FPA as easier to use and more useful than IFPUG FPA. Besides, the participants are more likely to use Extended FPA than IFPUG FPA in calculating software security cost.

Acknowledgments. The researchers want to thank all the participants who were taking part in the experiment. Thanks to University of Technology MARA for sponsoring this paper. Special gratitude to Kartini Rashid and Mohd Faisal Ibrahim who were willing to review this paper.

References

1. Albrecht, A.J.: Measuring Application Development Productivity. In: Proceedings of IBM Application Development Symposium, pp. 83–92 (1979)
2. Come back function point analysis (modernised) all is forgiven, http://www.measuresw.com
3. Jones, T.C.: Estimating Software Costs. McGraw-Hill, United States of America (1998)
4. Symons, C.: Software Sizing and Estimating – Mark II FPA. Wiley Series in Software Engineering. Wiley, United Kingdom (1991)
5. Whitmire, S. A.: 3D Function Points: Specific and Real-Time Extensions of Function Points. In: Proceedings of the Pacific Northwest Software Quality Conference (1992)
6. Pierre, D., St., M.M., Abran, A., Desharnais, J.M., Bourque, P.: Full Function Points: Counting Practices Manual. Software Engineering Management Research Laboratory and Software Engineering Laboratory in Applied Metrics Technical Report 1997-04 (1997)

7. Symons, C., Lesterhuis, A.: The COSMIC Functional Size Measurement Method Version 3.0.1 – Measurement Manual. Common Software Measurement International Consortium: United Kingdom (2009)
8. Mercuri, R.T.: Analyzing Security Cost. Communications of the ACM 46(6), 15–18 (2003)
9. McGraw, G., Routh, J.: Software in Security: Measuring Software Security. InformIT (2009)
10. Jari, R.: Contracting over the Quality Aspect of Security in Software Product Markets. In: Proc. of 2nd ACM workshop on Quality of Protection, pp. 19–26 (2006)
11. Boehm, B. W., Abts, C., Brown, A.W., Chulani, S., Clark, B.K., Horowitz, E., Modachy, R., Reifer, D., Steece, B.: Software Cost Estimation with COCOMO II. Prentice Hall, Inc., New Jersey (2000)
12. Reifer, D.J., Boehm, B.W., Gangadharan, M.: Estimating the cost of Security for COTS Software. Technical Report USC-CSE-2003-502 (2003)
13. Colbert, E., Wu, D., Chen, Y., Boehm, B.: Cost Estimation for Secure Software and Systems 2006 Project Update. University of Southern California, Center of Software Engineering, United States of America (2006)
14. Colbert, E., Boehm, B.: Cost Estimation for Secure Software & Systems. Paper presented at the ISPA/SCEA 2008 Joint International Conference, The Netherlands (2008)
15. Wu, D., Yang, Y.: Towards an Approach for Security Risk Analysis in COTS Based Development. In: Proceedings of Software Process Workshop on Software Process Simulation, Center for Software Engineering. University of Southern California, Los Angeles (2006)
16. Abdullah, N.A.S., Selamat, M.H., Abdullah, R., Jaafar, A.: Potential Security Factors in Software Cost Estimation. In: Proceedings of the International Symposium on Information Technology (ITSim 2008), vol. 3. IEEE, Kuala Lumpur (2008)
17. Abdullah, N.A.S., Selamat, M.H., Abdullah, R., Jaafar, A.: Software Security Characteristics for Function Point Analysis. In: Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, Hong Kong (2009)
18. Abdullah, N.A.S., Selamat, M.H., Abdullah, R., Jaafar, A.: Security Risks Formulation in Software Cost Estimation. In: Proceedings of the Malaysian Software Engineering Conference (MySEC 2008). IEEE, Terengganu (2008)
19. Davis, J.: Internet Banking. Paper presented at the I-Hack. University of Technology MARA, Malaysia (2007)
20. McGraw, G.: Software Security: Building Security, United States of America. Addison-Wesley Software Security Series. Pearson Education Inc., London (2006)
21. Department of Education, Information Technology Security: Information Technology Security Cost Estimation Guide (2002)
22. CCRA Working Group: Common Criteria for Information Technology Security Evaluation Version 3.1 (2009)
23. Wiesmann, A., Stock, A., Curphey, M., Stirbei, R.: A Guide to Building Secure Web and Web Services. In: Black Hat 2nd Edition; The Open Web Application Security Project (2005)
24. IT Governance Institute: Control Objectives for Information and related Technology (COBIT 4.1): Framework, Control Objectives, Management Guidelines, and Maturity Models (2007)

25. Abdullah, N.A.S., Abdullah, R., Selamat, M.H., Jaafar, A., Jusoff, K.: Estimating Software Cost with Security Risk Potential. International Journal of Economics and Finance, Canadian Center of Science and Education 1(1), 183–192 (2009)
26. Bond, T.G., Fox, C.M.: Applying the Rasch Model: fundamental measurement in the Human Sciences, 2nd edn. Lawrence Erlbaum Associates, Inc., New Jersey (2007)
27. Abdullah, N. A. S., Selamat, M. H., Abdullah, R., Jaafar, A.: Validation of Security Awareness in Software Development by using RASCH Measurement. Paper presented in the Pacific Rim Objective Measurement Symposium, Hong Kong (2009)
28. Abdullah, N.A.S., Selamat, M.H., Abdullah, R., Jaafar, A.: Extended Function Point Analysis Prototype - With Security Costing Estimation. In: Proceedings of the International Symposium on Information Technology (ITSim), Kuala Lumpur, Malaysia, June 15- 17, IEEE, Malaysia (2010)
29. Davis, F.D.: Perceived Usefulness. Perceived Ease of Use and User Acceptance of Information Technology. MIS Quarterly 13(3) (1989)
30. Fishbein, M., Ajzen, I.: Beliefs, Attitude, Intention and Behavior. An Introduction to Theory and Research, Reading, MA (1995)
31. Moody, D.L.: The Method Evaluation Model: A Theoretical Model for Validating Information Systems Design Methods. Unpublished paper in School of Business Systems, Monash University, Melbourne, Australia (2003)
32. Rescher, N.: The Primacy of Practice. Basil Blackwel, Oxford (1973)
33. Moody, D.L.: Comparative evaluation of large data model representation methods: The analyst's perspective. In: Spaccapietra, S., March, S.T., Kambayashi, Y. (eds.) ER 2002. LNCS, vol. 2503, p. 214. Springer, Heidelberg (2002)
34. ISO/IEC 14143-3-Information technology-Software measurement-Functional size measurement-Part 3: Verification of functional sizemeasurement methods (2003)
35. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers, Dordrecht (2000)
36. Abrahao, S., Poels, G., Pastor, O.: Assessing the Accuracy and Reproducibility of FSM Methods through Experimentation. In: 3rd International ACM-IEEE International Symposium on Empirical Software Engineering (ISESE 2004), Redondo Beach CA, USA (August 2004); ISBN 0-7695-2165-7
37. Leedy, P.D., Ormrod, J.E.: Practical Research: Planning and design, 8th edn. Pearson Educational International and Prentice Hall, New Jersey (2005)

Analyzing Framework for Alternate Execution of Workflows for Different Types of Intrusion Threats

Sohail Safdar and Mohd Fadzil Hassan

Department of Computer and Information Sciences
Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Tronoh, Perak, Malaysia
sagi_636@yahoo.com, mfadzil_hassan@petronas.com.my

Abstract. The main objective of the research is to conduct an analysis of the framework for alternate execution of workflows under intrusion threat with respect to different types of threats. Framework for alternate execution of workflows under threat makes the system available to the end user no matter if it is under attack by some intrusion threat. The assessment is required to be made for the framework in consideration in terms of what types of threats and how many types of threats for which it may work. For this purpose, 34 different types of threats as described by SOPHOS have been considered. Firstly the types of threats have been categorized based on their goals. Then for each category, framework in consideration is assessed. On the basis of that assessment it is analyzed for what types of threats, the framework can be enabled completely and partially. The number of threats for which the framework is enabled completely is also found. Based on the analysis, the recommendations have been made for possible extensions in the framework where it is enabled partially.

Keywords: Alternate execution; Data Hibernation; SOPHOS; Workflows.

1 Introduction

Modern world has emerged to be developing with their fast growing economy. Information technology has brought major advancements in almost every field of life. The pace of performing any task is far more than that of the early times. In this era of high market competition, business enterprises are focused to provide high quality of services to their consumers. Acquiring maximum consumer satisfaction is the prime goal of every business enterprise. Business enterprises are relying heavily on IT infrastructures to strengthen their business processes.

Workflow Management System (WFMS) is very hot area in research as they are managing, monitoring and controlling business workflows for attaining certain business goals. Workflow is defined for any business process that needs to be carried out at a certain time. Workflows are software processes that have a certain flow of execution based on the flow of business process. Business processes may vary from standalone application process to the online web based service depending on the nature of the business. Optimization of business processes and workflows is the continuous process that depends on thorough research. Various efforts have been made to increase the performance of the workflows, making workflows robust, effective design of workflows and their security.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 361–375, 2011.

As discussed before that business enterprises focus a lot on providing quality of services to their consumers to acquire their maximum satisfaction. For this purpose, consumers should be facilitated with the high performance, robust and 100% available software systems to carry out their desired transactions in efficient fashion. No matter how high the system performance is, when there is an intrusion attack to the system, it has to stop executing. All of the executing transactions should be stopped and rollback to avoid inconsistency in the system state. The continuity of the desired workflow cannot be promised until the system is recovered from the intrusion threat and becomes up again to be accessed. Hence, whenever there is an intrusion attack to the system, the system becomes temporarily unavailable and goes in to wait state until it is recovered from that threat. This temporary unavailability of the system in case of threat scenarios may cause consumer dissatisfaction and they might lose trust on the system as well in some cases. It is therefore desirable to have the system that can remain not only available in the state when it is intruded by a threat but also continue its execution robustly and in a secure fashion. Current research is the extension of the research that addresses the issue of making secure availability of the system when it is attacked by an intrusion threat.

Framework for alternate execution of workflows is proposed with an objective to provide the secure availability of the workflows when the system is under an intrusion attack. This alternate execution framework is aimed to provide the availability of the system in such a way that the integrity of the data should remain intact and the access to the resources should be made only by the safe users. The framework comprises of two major components. One is Special Authentication whereas the second is Data Hibernation. Special authentication is the unconventional and more secure way of authenticating a user so that user may be declared as safe user to access the system resources. The concept of Data hibernation is extracted from the hibernation in animal in which it hibernates itself for a certain period of time due to environmental constraints. Data hibernation is the transferring of data from its original source to some alternative source so that it may be available from the place other than that being under attack.

The framework in its basic implication is useful. However the question arises whether the framework for alternate execution can be useful in all types of threat attacks. If the threat attack makes some data dirty, so it becomes infeasible to shift that data to alternate source. Hence the problem is to find out for what types of threats the alternate execution framework is useful in its current form and for what types of threats it should be extended to handle them appropriately. The main objective of the current research is to analyze the framework for alternate execution of workflows under different types of threats. The 34 types of threats that have been described by SOPHOS [3] are considered for analysis. First the types of the threats have been categorized by their potential goals and then the analysis of framework in consideration has been performed. Based on the analysis, it is concluded for what types of threats the framework can be applied as it is, and for what other types of threats it requires an extension. The required extension to the existing framework is recommended and its pros and cons are also discussed.

The rest of the paper is organized five further sections. Section 2 enlightens the background of the research. Section 3 describes the framework for alternate execution of workflows under threat in detail. Section 4 categorizes and resolves different types of threats as described by SOPHOS [3]. Section 5 provides the analysis for how many types of threats, the framework is workable and recommendations that have been made followed by the conclusion and references.

2 Background

Researchers have been putting their efforts to enhance workflows in the domain of their representations, adaptive behavior, performance, management issues, security and self healing. Current research focuses merely to the system availability in case of intrusion threat along with putting much concern for the security and integrity of the data. As there is no significant work done in making system available in case of intrusion threat so far but various efforts have been made proposing different approaches for intrusion detection and system recovery in workflows. One of an approach is the use of Multi-Version Objects [1] [7] [25] to recover the system to its consistent state. In Multi-Version Objects based approach, the data object that becomes dirty due to the intrusion attack should be replaced with the clean replica version of the respective object to recover the system. The system works by having more than one version of each data object that is involved in the transaction. The dirty version of an object is replaced with the clean and consistent version for recovery purpose when system suffered from an intrusion attack. Another technique that is Trace back recovery [1] [8] [25] is based on Flow-Back Recovery Model [1] [22] that recovers the system by tracking the traces of the flow of execution of workflow. One of the techniques detects the intrusion by the workflow specification using independent Intrusion Detection System (IDS). Then "Attack Tree Model" [7] [9] [25] is drawn to analyze and describe the major attack goals and their sub goals. The system recovery is then done by dynamic regeneration of workflow specification. Architecture such as BPEL (Business Process Enterprise Language) Engine and Prolog Engine for intelligence is used to regenerate the workflow specification dynamically [9]. MANET [10] on the other hand provides mobile services, workflow modeler and policy decision point to regenerate the workflow specification [1] [10] [25]. Use of workflow layer as a non intrusive approach to detect the attack in the system is proposed for surviving in the cyber environment [11]. One of the proposed facts is threat agent causes threats that lead to vulnerability [12] [25]. The risks caused by those vulnerabilities can be reduced by using a safe guard to protect an asset [12]. Other approaches are also there such as Do-It-All-Up-Front, All or Nothing, Threat Modeling and Big Bang etc. to provide security on web [1] [13] [14]. All transactions done by malicious users is undo and cleaning of dirty data is done to recover the system [17]. There are number of algorithms applied to recover the system based on their dependencies information [23]. Potential failures in workflows can also be studied and becomes a source of possible recovery mechanism [1] [20] [25]. Handling the problems associated to the recovery and rollback mechanisms in distributed environment is also studied [21]. Furthermore there is a work related concurrency control in databases and its transaction is also done [18] [19]. It may be noted that all recovery techniques and approaches discussed above works only if the system goes offline while it is under an intrusion attack. In case of intrusion attack, there is a need to undo all the currently executing activities and system goes offline until it is being recovered by applying some technique. The transactions and activities should be redone only once the system is recovered successfully and becomes online again. Therefore the system remains in a wait state and becomes temporarily unavailable the whole time when it gets under attack till it is recovered. For attaining maximum consumer satisfaction, business enterprises desire the availability of their systems in all conditions even if it is an intrusion attack [1] [2] [25] [26]. Making system availability in the situation when it is under attack by some intrusion threat requires ensuring the security and integrity of the data along with the controlled access to the resources. But it is impossible to keep the system available in its active state as the intrusion may result stealing of data or may corrupt the data physically. To avoid this loss of data, the system should go offline and cease all its activities. The goal of making system available under these conditions can be achieved with the help of framework for alternate execution of workflows under intrusion threat [1]. During intrusion attack, the actual system becomes offline but the continuation of the services is provided using an alternate execution path using an alternate data sources. It consists of two major components, special Authentication and data hibernation [1] [2] [25] [26] [27]. The special authentication is meant to provide the access of the resources only to the safe users and avoid malicious accesses. The authentication may be done by using biometric systems [6], hardware based RFID [4], graphical passwords [5], multi-mode [16], encrypted [15] or textual such as two dimensional passwords [2] [26]. Applying two dimensional passwords in this scenario is proposed as they are not so expensive compared to all other techniques provided they possess tremendous amount of strength as well [2] [26]. The strength is calculated in terms of bit entropy [24]. Moreover, the small and medium scale organizations that cannot afford expensive solution for authentication can get maximum use out of it [2]. Simultaneously with the authentication process, the data has to be transfer to the alternate sources applying data hibernation [27]. Framework for alternate execution of workflows makes it possible for them to continue their execution robustly even in the vulnerable state. The details of the framework for alternate execution of workflows under intrusion threat are given in the next section.

3 Framework for Alternate Execution of Workflows

Framework for Alternate execution of workflows is proposed for to make workflows execute robustly even in the scenario when they are under some intrusion attack and the software system remains available to the end user or consumers. System availability is a great challenge when the workflow is under an intrusion threat. The data integrity is challenged and authentication been misused by malicious users. To avoid this security loss, the system is supposed to get offline immediately once the intrusion threat is detected. Framework for alternate execution come in to play here and provides and alternate ground of execution, so that the system becomes available to the consumers and end users without inconvenience. However the actual system setup needs to be recovered in order run the system back to its original pattern. During the system recovery the consumers can perform their tasks conveniently to the system executing at alternate path. The framework for alternate execution of workflows is aimed for the software systems that come under the domain of LAN based systems, Web-based online systems, and other distributed software systems. The framework in consideration consists of two major components, one is Special Authentication and the other is Data Hibernation. Special authentication aims to provide the strong authentication to avoid any malicious access to the alternatively executing system whereas the data hibernation aims to transfer data from the original data source to more secure alternate data source i.e. dimensions. Dimensions are specially designed so that data integrity cannot be challenged at any cost. These are small subsets of original database. Using dimensions allows you to access data from more than one location making it difficult for hackers to hack the information completely in a meaningful way. Special authentication can be done using two dimensional passwords so that maximum strength can be achieved by using textual passwords. Another aim of using two dimensional passwords is that the small and medium scale organizations can afford the solution and becomes capable of using framework for alternate execution of workflows during intrusion threat. The components and working of the framework in consideration is shown in figure 1[1].

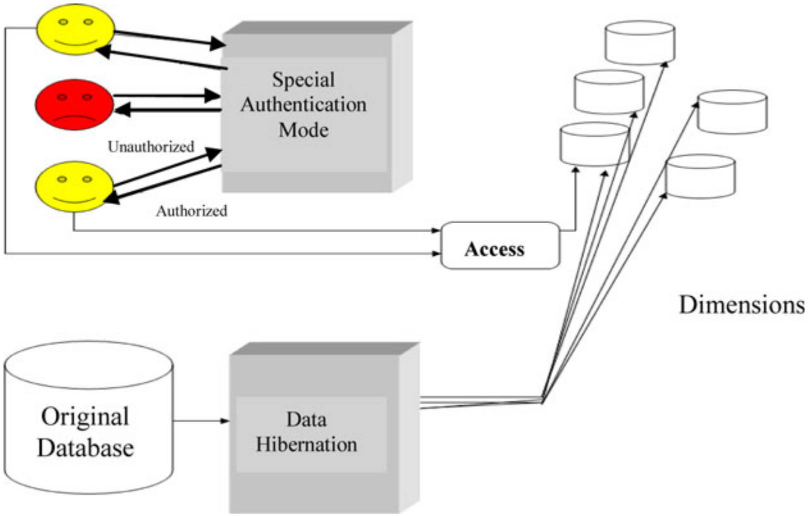


Fig. 1. Alternate Execution Framework Architecture

When the intrusion threat is detected than the system goes offline, disconnecting all users considering them as unsafe users and data hibernation started to shift the data to the dimensions. Users have to re-login using special authentication module by applying their two dimensional passwords. Once the authentication is successful then the users have to change their linear passwords. User can now access to their data and perform their transactions as the data is hibernated parallel to their authentication process. Special authentication is one time process and from the next time the linear passwords can be used to access an account.

4 Categorization of Types of Threats

It is very important to identify all those threats for which the framework for alternate execution of workflows under intrusion threat can be enabled completely or partially.

For this purpose, the 34 different types of threats as defined by SOPHOS [3] are considered. Before directly moving to analyze the framework, it is required that these types of threats are categorized based on the goals that they aimed to achieve. Therefore, few threat goals oriented categories are defined as shown in Table 1. It is believed that all considered 34 types of threats lies in to these categories based on the goals that they aimed to achieve.

Table 1. Categories of Threats Based on Goals

Sr #	Threats Category based on Goals	Description
1	Advertisement (Adverts)	Threats that are meant for advertisement using the local resources of the users without their consent.
2	Anonymous Messages (Msgs)	Threats that are meant to pass personal messages using the local resources of the users without their consent.
3	Information Stealers (Stealers)	Stealing useful information by intruding but do not corrupt or modify the data.
4	Modifiers (Mod)	Threats that intrude for modification, addition and updating the data maliciously hence corrupting useful information of an organization.
5	Crash	Destroys the system by crashing the services, i.e. permanent inaccessibility of system.
6	Prohibit	Controls and prohibits the access to certain useful activities and resources.
7	Divert	Controls and redirects the access to certain activity and resource to some other malicious activity or resource.
8	Confidentiality Compromise (CC)	Compromising privacy of users in terms of personal information or activities performed by the user.
9	Halt	Means the user cannot get out of some activity even if it is completed successfully.
10	Change Setting (CSet)	Threats that change the computer settings and then any other stealing threat can steal information by attacking the system.
11	Hider	Most dangerous threats that can hide any type threat programs and processes.
12	Forgery	To deceive someone to earn money by forged information.

Table 1 shows these categories along with their description. Long named categories are aimed to be used by the abbreviations assigned to them as mentioned in table 1. Once the categories are defined, it is required to describe all 34 types of threats as defined by SOPHOS [3] in terms of attacking the organizations' data. Based on the description and their goals, these types of threats should be categorized based on table 1. Table 2, describes all 34 types of threats, categorize them along with their impact. The impact may be high (H), medium (M) or low (L) based on their description of the way of attacking and damaging.

Table 2. Description and Categorization of Types of Threats

Sr #	Threat Type	Description/purpose	Category Assigned	Impact
1	Adware	Main purpose is to download the advertisements without the consent of user. Slowdown the connectivity. May be used as downloading other's browser information and sending local information to other's browser.	Adverts, Stealers, Msgs	H
2	Backdoor Trojans	It adds to the computer startup routine. It can run programs to infect computer, access files and steal information, modify information and upload programs and files.	Stealers, Mod	H
3	Blue-jacking	Sending Anonymous messages to other phones and laptops using Bluetooth. It does not steal information and does not take control of the device. Just used for unwanted messages.	Adverts	L
4	Blue-snarfing	It connects to the Bluetooth device without the consent or approval. Steal useful information.	Stealers	H
5	Boot Sector Viruses	Modifies the startup program so that computer runs the virus when boot up, hence crashes and faces loss of data. Virus can infect the software system services and crash it. These are rarely encountered today as being old type viruses.	Crash	M
6	Browser Hijackers	Hijacks the browse and forcefully redirects the users to visit their targeted website. The redirects may be aimed for marketing of some website or may be to prohibit the users to visit a certain online system.	Adverts, Prohibit, Divert	H
7	Chain Letters	Depends on anonymous users instead of computer code. Propagate messages, pranks, jokes, petitions and false claims etc. Usually propagated via emails or sms. Aimed to waste time. Do not harm security.	Msgs, Adverts, Forgery	L
8	Cookies	These are not the threat to information of an organization but can compromise the confidentiality of the users. Usually not severely harmful but cannot be ignored if confidentiality of the user is an issue by the business process. Cookies itself is not harmful but can be accessed by anonymous users and responsibly for compromising confidentiality.	CC	M

Table 2. (*continued*)

Sr #	Threat Type	Description/purpose	Category Assigned	Impact
9	Denial-of-Service (DoS)	Aims to deny the services to the potential users. This is done by overpopulating the request buffer so that services required by the potential users become inaccessible. The methods may be sending large number of bogus requests for the connection or using IP ping message from the victim's computer so that it receives a huge amount of responses. No information stolen or compromised.	Prohibit	H
10	Dialers	Limited to Dial-up users only. Usually diverts users to premium number connections to charge them more. Do not steal or modify information. Installs itself with and without the consent of the users.	Divert	L
11	Document Viruses	Spread through Macros associated to the documents. Can steal and Modify information. Can be as severe as to crash the targeted system. Disabling macros is suggested.	Stealers, Mod, Crash	H
12	Email Viruses	Distribute automatically through emails. Can be program scripts or bogus messages. Can access the system and steal data. Risks the security. Major concern is to increase traffic that is accessing the system.	Stealers, Mod, Crash	H
13	Internet Worms	Use communication between the computers to propagate them. Aimed to attack the systems in terms of overpopulating the services to prohibit the access i.e. DoS and may also crash the system.	Prohibit, Crash	H
14	Mobile Phone Viruses	Uses mobile phones to spread. Cannot directly harm the system and cannot steal or modify the information. Cannot be spread easily as the operation systems are different in phones.	Msgs	L
15	Mouse-trapping	It forces you not to leave a certain web page. The page may contain virus, spyware or Trojans. Does not directly steal or modify information. May be aimed for advertisements or forgery. Cannot work alone, originated as a result of diversion.	Adverts, Prohibit, Halt, Forgery	M
16	Obfuscated Spam	Attempt to fool anti spam by modifying the spam keywords so that they may not be detected. Used mostly for Advertisements purpose. Use of spaces, HTML codes and other way of hiding texts. Does not harm the system's security.	Adverts, Forgery	L

Table 2. (*continued*)

Sr #	Threat Type	Description/purpose	Category Assigned	Impact
17	Page-Jacking	It makes the replicas of the reputable online system's pages. These replicas are then used to steal the personal and other information from the users. The company's information is also become at stake using this technique. Company may lose consumer trust also. It is originated as a result of diverting to false webpage.	Divert, Stealers	H
18	Palmtop Viruses	Usually spread by palm tops when they are connected to the computer. The virus remains harmless but installs and then attacks the system when it is transferred to the computer. But it can easily be avoided if the portable devices are not connected to the important computer that runs a desired software system.	Stealers, Mod, crash	L
19	Parasitic Viruses	Old type of viruses but still can harm. Fool the operating system by posing them as potential program to get the same access right. Installs and runs to make changes in the computer settings. In this way they are prone to leak information from the computer, may be the important information of businesses.	CSet	M
20	Pharming	Diverts to the bogus copy of the legitimate site and allow stealing of important business information along with the personal information.	Divert, Stealers	H
21	Phishing	Tricking users using bogus websites and emails to feed in important confidential information. Sometimes part of the bogus websites is enabled in the legitimate website to steal information.	Divert, Stealers	H
22	Potentially Unwanted Applications (PUAs)	These are programs that may be used for advertising and are not malicious. But these are not suitable for company networks and websites as they may open the loop holes for the malicious programs to attack.	Adverts, Crash, Msgs	M
23	Ransomware	The one that denies access to the required resource until the ransom is paid. It may steal or may not harm the system but makes a bluff that system or service may crash in three day etc. if ransom is not paid. Aim is to earn money by acquiring control, even if the data cannot be steal. Asymmetric encryption is suggested to use at server end to avoid this.	Prohibit, Stealers, Msgs, Adverts, Forgery	H

Table 2. (*continued*)

Sr #	Threat Type	Description/purpose	Category Assigned	Impact
24	Rootkit	It is used to hide the program and processes running on the computer. It hides the running viruses that may lead to information loss, stealing information and crashing programs or processes. Cannot be detected easily.	Hider	H
25	Share Price Scams	Share prices are falsely been published through these scams. Artificial rising market stats are used and after selling share at profit, the price collapses. Do not directly works from the company's software system.	Forgery	L
26	Spam	Used for marketing and advertisements through emails and messages. Some spam propagates viruses that may be used to filling up the company's database with false information and using its bandwidth. At times employers may hold responsible for something they haven't done.	Adverts, Mod	M
27	Spear Phishing	Well targeted way of persuading company people to reveal their usernames and passwords. The query seems to be generated from trusted department and leads to information stealing.	Stealers	H
28	Spoofing	Uses wrong sender information to steal the important information from the victim user. Results in the important information loss of the company.	Stealers	H
29	Spyware	They are not installed to the user's computer but observe the user's activities when he/she visited some site. It is software that shares the important information of the company and its users to the advertisers and hackers.	Stealers	H
30	Trojan Horses	Exposed as legitimate programs but carry out unwanted agenda of stealing information and enabling viruses. Works hand-in hand with viruses, therefore may prone to infect data as well.	Stealers, Mod, Prohibit, Crash.	H
31	Viruses	These are the programs that run on the computer before they can actually attack their target software systems or applications. They can steal information, modify information and add false information. Dangerous along with Trojan horses.	Stealers, Mod, Crash.	H
32	Virus Hoaxes	Reports of non-existent viruses. Overload mail server for the victim company and it cannot read or write mails. As they are not viruses so their detection are difficult. They may lead to crash the mailing server as well.	Crash, Prohibit.	H

Table 2. (*continued*)

Sr #	Threat Type	Description/purpose	Category Assigned	Impact
33	Voice Phishing	Depends on telephony conversation to make frauds to the people.	Forgery	H
34	Zombies	It is a computer that is controlled remotely without the knowledge of the user of company. Trojans and viruses open backdoor and that information is used to control that user's computing resources. Hence that node becomes zombie.	Stealers, Forgery, Crash, Prohibit, Divert	H

It may be observed from table two that one type of threat can have multiple goals, therefore can be assigned multiple categories whereas one threat category can also have multiple threat type assigned. Hence the relation between threat types to categories is termed as Many-Many relation by nature.

5 Analysis of the Framework w.r.t. Types of Threats

Based on the categorization of all types of threats, it is now possible to assess how many types of threats and what types of threats are there for which the framework for alternate execution of workflows under intrusion threats can be enabled. The analysis consists of two steps. In first step, the framework's capability of handling all the defined categories is assessed. The capability of the framework is assessed on these parameters i.e. Completely, Partial, Not Capable and Not Required. These parameters suggests whether the framework can be applied as completely, partially or is not required to provide alternate path of execution to the workflows when they encountered certain types of intrusion threats. In the second step, the types of threats for which the framework can be enabled completely, partially or not required can be found out based on the category analysis in the first step. Table 3 shows the assessment of the categories for which the framework for alternate execution of workflows can work completely or partially or is not required.

Table 3. Assessment of the Framework based on the Threat Categories

Sr #	Threats Category	Description of Application	Application
1	Advertisement (Adverts)	If detected, no need to apply framework but to handle using conventional anti spam software.	Not Required
2	Anonymous Messages (Msgs)	If detected, no need to apply framework but to handle using proper investigation.	Not Required
3	Information Stealers (Stealers)	If detected, framework can be applied completely i.e. to provide alternate access to the user with specialized authentication with an alternate data source. Information Stealers cannot access the alternate sources as being unaware.	Complete

Table 3. (*continued*)

Sr #	Threats Category	Description of Application	Application
4	Modifiers (Mod)	If detected, framework should be applied completely with an additional component that can handle to reset the corrupted to the last consistent state before it is being hibernated, otherwise it cannot work in this scenario.	Not Capable
5	Crash	If detected, framework can be applied completely.	Complete
6	Prohibit	If detected, framework can be applied completely.	Complete
7	Divert	If detected, framework can be applied completely.	Complete
8	Confidentiality Compromise (CC)	If detected, framework can be applied completely.	Complete
9	Halt	If detected, framework can be applied completely.	Complete
10	Change Setting(CSet)	If detected, Analyze the change and then framework may or may not be applied.	Complete
11	Hider	If detected, framework can be applied completely but for modification operations it should be applied with extension.	Partial
12	Forgery	If detected, no need to apply framework but to handle using conventional anti spam software.	Not Required

On the basis of the assessment in table 3, the following table 4 provides the types of threats for which the framework in consideration can be enabled completely, partially or is not required.

Table 4. Analysis of Framework for 34 Types of Threats

Not Required for Threats	Partially Enabled for Threats	Completely Enabled for threats
1. Bluejacking	1. Backdoor Trojans	1. Adware
2. Chain Letters	2. Document Viruses	2. Bluesnarfing
3. Mobile Phone Viruses	3. Email Viruses	3. Boot Sector Viruses
4. Obfuscated Spam	4. Palmtop Viruses	4. Browser Hijackers
5. Share Price Scams	5. Rootkit	5. Cookies
6. Voice Phishing	6. Spam	6. Dialers
	7. Trojan Horses	7. Denial-of-Service (DoS)
	8. Viruses	8. Internet Worms
		9. Mousetrapping
		10. Page-Jacking
		11. Parasitic Viruses
		12. Pharming
		13. Phishing
		14. Potentially Unwanted Applications (PUAs)
		15. Ransomware
		16. Spear Phishing
		17. Spoofing
		18. Spyware
		19. Virus Hoaxes
		20. Zombies

It may be noted from table 4 that there is no threat type for which the framework is unable to operate, however the types where the framework enabled as partial is because the existing framework requires an additional component to clean the data from the malicious modifications made by the threat. Table 1, defines the categories of the threats based on the goals. Table 2, categorizes the considered types of threats based on table 1. Once the types of threats are categorized in table 2, only then it is possible to assess the framework in terms of how it is working for different types of threats because framework's working is directly addressing to the goals of the threats as mentioned in table 3. Table 4 gives the analysis of the framework for 34 different types of threats whether it is workable or not.

5.1 Outcomes and Recommendations Based on the Findings

Table 5 provides the number of threats for which the framework in consideration works as complete, partial or not required. Even if the framework is not enabled completely in some cases but it is capable of working in that scenario partially rather then it being incapable.

Table 5. Analysis in Terms of Number of Threats

Sr #	Enabling Level		Total Types of Threats
1	Not Required		6
2	Partially Enabled		8
3	Completely Enabled		20
4	Incapable		0
		Total	34

From the above analysis, it can be seen that framework for alternate execution of workflows under intrusion threat can work for majority of threats scenarios. The only threat category for which the framework requires an extension to deal with is the one that can modify and update the data maliciously. For handling such threats, the framework in consideration should have a component that is meant to clean the data and works well before the data hibernation starts. The problem arises here is that, the data is supposed to be clean before it is hibernated, therefore data hibernation step may be delayed which is not acceptable by this framework to give the desirable results. Hence it is recommended to apply the policy engine to decide how the most active data should be clean in time to make the system available and workable as soon as possible. Based on the policy made by the policy engine that depends on the execution state of the system at that particular time when it is being attacked, the quick data cleaning step should be introduced followed by the data hibernation. The data hibernation may also work in to two steps i.e. quick hibernation of maximum clean data and then later the rest of the clean data should be transferred with the time. From table 3, table 4 and table 5, it is clearly seen that the considered framework is applicable to most of the intrusion threats in its current structure. For the rest of the few types of threats, framework requires an extension to become completely workable. The extension of the framework contains its own pros and cons. It may cause certain delays in data hibernation process too, that should be handled smartly.

6 Conclusion and Future Work

Framework for alternate execution of workflows under intrusion threats is a pioneer step in making the system available when it is under attack. The results acquired in this research support the argument that this framework is useful in most of the types of intrusion threat attacks. It deals with the variety of threats and can be enabled to provide the end user with the robust execution of their desired tasks. Businesses take the advantages by using this framework in terms of robust execution of their workflows and consumer satisfaction as well as trust in their systems. However there are few types of intrusion threats for which the framework cannot be applied completely. Hence in future, if the possible extensions are applied to the framework than it becomes equally useful for all types of threats. The extension has to be designed in such a way as it will not cost excess time in terms of unavailability of the system provided the security remained intact while the system is made workable under intrusion threat.

References

1. Safdar, S., Hassan, M.F., Qureshi, M.A., Akbar, R.: Framework for Alternate Execution of workflows under threat. In: 2nd International Conference on Communication Software and Networks ICCSN, Singapore (2010)
2. Safdar, S., Hassan, M.F.: Moving Towards Two Dimensional Passwords. In: International Symposium on Information Technology ITSIM, Malaysia (2010)
3. SOPHOS: a to z of Computer Secutity Threats SOPHOS (2006) http://security.ucdavis.edu/sophos_atoz.pdf(retrieved on January 17, 2011)
4. Yang, D., Yang, B.: A New Password Authentication Scheme Using Fuzzy Extractor with Smart Card. In: 2009 International Conference on Computational Intelligence and Security, pp. 278–282 (2009)
5. Oorschot, P.C.V., Thorpe, J.: On Predictive Models and User-Drawn Graphical Passwords. ACM Transactions on Information and System Security 10(4), article 17 (2008)
6. ChunLei, L., YunHong, W., LiNing, L.: A Biometric Templates Secure Transmission Method Based on Bi-layer Watermarking and PKI. In: 2009 International Conference on Multimedia Information Networking and Security, China (2009)
7. Meng, Y., Peng, L., Wanyu, Z.: Multi-Version Attack Recovery for Workflow Systems. In: 19th Annual Computer Security Applications Conference ACSAC,1063-9527/03. IEEE, Los Alamitos (2003)
8. Meng, Y., Peng, L., Wanyu, Z.: Self-Healing Workflow Systems under Attacks. In: 24th International Conference on Distributed Computing Systems ICDCS, 1063-6927/04. IEEE, Los Alamitos (2004)
9. Fung, C.K., Hung, P.C.K.: System Recovery through Dynamic Regeneration of Workflow Specification. In: Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing ISORC, 0-7695-2356-0/05. IEEE, Los Alamitos (2005)

10. Fung, C.K., Hung, P.C.K., Kearns, W.M., Uczekaj, S.A.: Dynamic Regeneration of Workflow Specification with Access Control Requirements in MANET. In: International Conference on Web Services ICWS, 0-7695-2669-1/06. IEEE, Los Alamitos (2006)
11. Xiao, K., Chen, N., Ren, S., Kwiat, K., Macalik, M.: A Workflow-based Non-intrusive Approach for Enhancing the Survivability of Critical Infrastructures in Cyber Environment. In: Third International Workshop on Software Engineering for Secure Systems SESS, 0-7695-2952-6/07. IEEE, Los Alamitos (2007)
12. Goluch, G., Ekelhart, A., Fenz, S., Jakoubi, S., Tjoa, S., Mück, T.: Integration of an Ontological Information Security Concept in Risk-Aware Business Process Management. In: 41st Hawaii International Conference on System Sciences. IEEE, Los Alamitos (2008)
13. Meier, J.D.: Web Application Security Engineering. IEEE Security Magazine,1540-7993/06, 16–24 (2006)
14. Virdell, M.: Business processes and workflow in the Web services world, `http://www.ibm.com/developerworks/webservices/library/ws-work.html`, IBM (2003) (retrieved on January 17, 2011)
15. Mitchell, S.: Encrypting Sensitive Data in a Database. In: MSDN Spotlight (2005)
16. Hsueh, S.: Database Encryption in SQL Server 2008. Enterprise Edition SQL Server Technical Article (2008)
17. Ammann, P., Jajodia, S., Liu, P.: Recovery from malicious transactions. IEEE Transactions on Knowledge and Data Engineering 14, 1167–1185 (2002)
18. Bernstein, P.A., Hadzilacos, V., Goodman, N.: Concurrency Control and Recovery in Database Systems. Addison-Wesley, Reading (1987)
19. Chrysanthis. P.: A framework for modeling and reasoning out extended transactions. PhD thesis, University of Massachusetts, Amherst, Amherst, Massachusetts (1991)
20. Eder, J., Liebhart, W.: Workflow Recovery. In: Conference on Cooperative Information Systems, pp. 124–134 (1996)
21. Gore, M.M., Ghosh, R.K.: Recovery in Distributed Extended Long-lived Transaction Models. In: 6th International Conference DataBase Systems for Advanced Applications, pp. 313–320 (1998)
22. Kiepuszewski, B., Muhlberger, R., Orlowska, M.: Flowback: Providing backward recovery for workflow systems. In: International Conference on Management of Data, pp. 555–557. ACM SIGMOD, New York (1998)
23. Lala, C., Panda, B.: Evaluating damage from cyber attacks. IEEE Transactions on Systems, Man and Cybernetics 31(4), 300–303 (2001)
24. Red Kestral, Random Password Strength (2004), `http://www.redkestrel.co.uk/Articles/RandomPasswordStrength.html` Red Kestral Consulting (retrieved on September14, 2009)
25. Safdar, S., Hassan, M.F., Qureshi, M.A., Akbar, R.: Biologically Inspired Execution Framework for Vulnerable Workflow Systems. International Journal of Computer Science and Information Security IJCSIS 6(1), 47–51 (2009)
26. Safdar, S., Hassan, M.F., Qureshi, M.A., Akbar, R.: Authentication Model Based on Reformation Mapping Method. In: International Conference on Information and Emerging Technologies ICIET IEEE, Pakistan (2010)
27. Safdar, S., Hassan, M.F., Qureshi, M.A., Akbar, R.: Data Hibernation Framework for Workflows under Intrusion Threat. In: 2011 IEEE Symposium on Computers and Informatics ISCI. IEEE, Los Alamitos (2011)

Taxonomy of C Overflow Vulnerabilities Attack

Nurul Haszeli Ahmad[1,2], Syed Ahmad Aljunid[2], and Jamalul-lail Ab Manan[1]

[1] MIMOS Berhad, TPM Bukit Jalil, 57000 Kuala Lumpur, Malaysia
[2] Faculty of Computer Sciences and Mathematics, UiTM,
Shah Alam 40000, Selangor, Malaysia
{haszeli.ahmad,jamalul.lail}@mimos.my, aljunid@tmsk.uitm.edu.my

Abstract. Various software vulnerabilities classifications have been constructed since the early 70s for correct understanding of vulnerabilities, and thus acts as a strong foundation to protect and prevent software from exploitation. However, despite all research efforts, exploitable vulnerabilities still exist in most major software, the most common still being C overflows vulnerabilities. C overflow vulnerabilities are the most frequent vulnerabilities to appear in various advisories with high impact or critical severity. Partially but significantly, this is due to the absence of a source code perspective taxonomy to address all types of C overflow vulnerabilities. Therefore, we propose this taxonomy, which also classifies the latest C overflow vulnerabilities into four new categories. We also describe ways to detect and overcome these vulnerabilities, and hence, acts as a valuable reference for developers and security analysts to identify potential security C loopholes so as to reduce or prevent exploitations altogether.

Keywords: Taxonomy, Classification, Buffer Overflow, Source Code Vulnerabilities, Software Security, Exploitable Vulnerability.

1 Introduction

Since the unintentional released of Morris Worm [1], experts have came out with many ways to protect system from overflow exploitation; either as a preventive measure or as runtime protection. There are more than 40 improvements or tools released, e.g. safe languages, extension of C, and safer C library to ensure software developed using C is secure and stable. [15], [16], and [17] have identified overflow vulnerabilities.

Unfortunately, C vulnerabilities exploitation is still a huge issue in software community [6], [7], and [8]. The classical overflow attack is still the most dominant vulnerability [9], [10], [11], and [19]. No doubt those earlier efforts have brought significant impact in reducing vulnerabilities and exploitation. However, improvements are still needed to eliminate or at least minimize C overflows vulnerabilities. Based on analysis on overflows vulnerabilities and evaluation on tools by [2], [3], [4], [5], [12], [13], and [14], we conclude the areas for improvement fall into three major categories; vulnerabilities understanding, analysis tool, and security implementation.

We limit our discussion to vulnerabilities understanding since accurate comprehension on the matter is a major step to improvement of security implementation

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 376–390, 2011.

and analysis tool, as shared by [18]. At present there is no taxonomy specifically addressing overflow vulnerabilities from C source code perspective. Therefore, we synthesize and propose a new taxonomy focusing on C overflow vulnerabilities that cover all vulnerabilities including four new types that have never been satisfactorily classified before. In addition, we also describe methods on detecting and overcoming these vulnerabilities.

Section 2 in this paper discuss briefly on type of taxonomy and some previous works. Section 3 views the proposed taxonomy while Section 4 summarizes the conclusion. Finally, Section 5 delineates future works on the taxonomy.

2 Previous Works

There are various taxonomies from different perspectives and scopes. Some shares the same issues and objectives such as analyzing tools and some were unique like verifying monitoring technique. All previous taxonomies have one main objective; to minimize exploitable software vulnerabilities.

[18], [19], [20], [21], [22], [23], [24], [25], [26], [28], and [29] presented general vulnerability taxonomies whereas [2], [3], [4], [32], and [33] constructed C vulnerabilities taxonomies focusing on C overflow vulnerabilities. Most of these taxonomies were later subsequently reviewed and analyzed further, as was done by [30]. Our work however focuses on C overflows vulnerabilities. As such, taxonomies that do not enclose or discuss C overflow vulnerabilities are ignored. Table 1 summarized our study on previous taxonomies focusing on scopes and types of C overflow vulnerabilities.

Table 1. Summary of Previous Vulnerabilities Taxonomies

Author	Scope	Type of C Overflows Vulnerabilities	Purpose/Objectives
Shahriar, H., Zulkernine, M. (2011)	Monitoring technique for vulnerabilities detection	Out-of-bound and few types of unsafe function.	Understanding the monitoring approaches
Alhazmi, O. H. et. al. (2006)	Cause and severity of software vulnerabilities at runtime.	Out-of-bound.	Develop testing plan and vulnerabilities projection
Tsipenyuk, K., et. al. (2005)	Common developer's mistake during coding in any programming language.	Out-of-bound, format string, pointer function, integer overflow, null termination, few memory function, and uninitialized variable.	Understanding common errors at coding stage.
Hansman, S., Hunt, R. (2005)	System security (software, network, and computer system)	General overflows.	Analysis and understanding of attacks
Moore, H. D. (2007)	Cause and Impact of overflows vulnerabilities	Out-of-bound, format string, and integer overflow.	Understanding of overflows vulnerabilities.

Table 1. (*continued*)

Author	Scope	Type of C Overflows Vulnerabilities	Purpose/Objectives
Sotirov, A. I. (2005)	C overflows vulnerabilities	Out-of-bound, format string, and integer overflow.	To evaluate static analysis tool and techniques.
Zhivich, M. A (2005)	C overflows vulnerabilities	Out-of-bound and format string	To evaluate dynamic analysis tool
Kratkiewicz, K. (2005)	C overflows vulnerabilities	Out-of-bound and format string	To evaluate static analysis tool
Zitser, M. (2003)	C overflows vulnerabilities	Out-of-bound and format string	To evaluate static analysis tool

While those past works on taxonomies have significant impact in reducing vulnerabilities and exploitation, renown security institutes and corporations [8], [10], and [11] continue issuing reports and advisories on C overflow vulnerabilities. Signifying breaches for exploratory discovery to aim for superior community comprehension of C overflows vulnerabilities. We ascertain four new types of overflow vulnerabilities necessitate classification as it is crucial.

3 Taxonomy of C Overflow Vulnerabilities Attack

We evaluate sets of vulnerabilities advisories and exploitations reports since 1988 until 2011. There are more than 50000 reported cases of C overflow vulnerabilities originating from five vulnerabilities databases and malware collection sites [9], [34], [6], [7], and [35].

From these reports, we classify types of C overflow vulnerabilities into ten categories. Four of them are new and still unclassified latest C overflow vulnerabilities. These are unsafe functions, return-into-libc, memory functions and variable type conversion. They have at least a medium level of severity, possibility to appear and exploited [6], [7], and [9]. The impact of exploitation with unsafe function is recognized as the most dangerous and outstanding [9], [34], [6], [7], and [35].

Figure 1 visualizes the new taxonomy of overflow vulnerability attack, organized in accordance to its severity, dominance, potential occurrence and impact. This taxonomy simplifies the understanding on implications of each types, their behavior and preventive mechanisms.

3.1 Unsafe Functions

Although unsafe functions has been exploited since 1988 [1], [15], 17], it is still relevant. More importantly, this well-known and well-documented inherent C security vulnerability is categorized as the most critical software vulnerabilities to continue to dominate C vulnerabilities report [6], [7], [35] and [39]. This implies there are software developers who are either ignorant, unaware, or simply bypass software security policies for prompt development [15], [16]. Below is a sample of unsafe functions vulnerability.

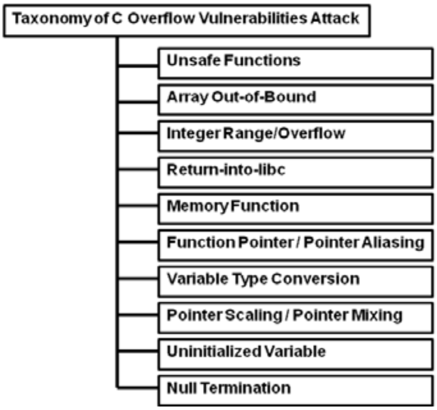


Fig. 1. Proposed Taxonomy for Overflow Vulnerabilities Attack in C

Part of a program showing *scanf()* vulnerability.

```
…
char str[20];
char str2[10];

scanf("%s",&str);
scanf("%s",&str2);
…
```

By supplying an input greater than the allocated size at the first *scanf()*, it automatically overflows the seconds variable and force the program to skip the second *scanf()*. This is one of many unsafe functions in C [12], [15], [16], [17], [36], [37] and [38]. Previous taxonomies classified few unsafe functions as Format String Attack, Read-Write, or Buffer-Overrun [2], [3], [15]. This is arguable since there are unsafe functions that do not implement formatting or require specifying index for reading or writing.

To prevent overflows via unsafe functions, one needs to check input variable before passing into any unsafe functions. Alternatively, there is C library safe functions that developers can use to avoid this type of overflow [17], [37].

3.2 Array Out-of-bound

Array Out-of-bound overflow can be triggered by misuse or improper handling of an array in a read or write operation, irrespective of it being above upper or below lower bound. A true sample is shown below.

A section from linux driver code in i810_dma.c contains the vulnerability [40], [41].

```
if(copy_from_user(&d, arg, sizeof(arg)))
      return -EFAULT;
```

```
if(d.idx > dma->buf_count)
      return -EINVAL;
buf = dma->buflist[d.idx]; //overflow if d.idx == -1
copy_from_user(buf_priv->virtual, d.address, d.used);
```

As shown in the above sample, when *d.idx* contains the value of -1, it will bypass the conditional statement which triggers overflow on the following statement. Array Out-of-bound overflows is easy to detect and prevent by monitoring all array processes and verifying whether the index is within the range specified; between zero to less than one from total array size.

3.3 Integer Range / Overflow

This type of overflow may occur due to miscalculation or wrong assumption in an arithmetic operation and is gaining its popularity in vulnerabilities databases [42], [43], [44]. The possibility of exploit is small, but the result of exploiting it is significantly dangerous [45].

This classification is derived from [26], [32], and [33]. The key difference is the removal of numerical conversion as one of the integer overflow type, and classifies it in a different category. This is due to its differences in behavior and code structure. Furthermore, the conversion errors are dependent on platform used to execute it. A true sample from [45] is shown below.

A fraction of C code contains Integer Range/Overflow vulnerability [45].

```
nresp = packet_get_int();
if (nresp > 0) {
response = xmalloc(nresp*sizeof(char*));
for   (i   =   0;   i   >   nresp;   i++)   response[i]   =
 packet_get_string(NULL);
}
```

As shown in the above code, if one able to inject input causing variable *nresp* to contain large integer, the operation *xmalloc(nresp*sizeof(char*))* will possibly trigger overflow, and later can be exploited [45]. It is difficult to detect as one needs to understand the logics and predict possible outcome from the arithmetic operation. As a result, this vulnerability tends to be left out undetected either by analysis tool or manual code review. This vulnerability can be avoided by simply restricting the possible input value before arithmetic operation took place.

3.4 Return-into-libc

Although it has been recognized as earlier as unsafe functions [84], it is yet to be appropriately classified. Many vulnerabilities databases rank its severity as high although the number of occurrence is low. It is difficult to detect since it can only appear during runtime and the code itself does not have specific characteristic to indicate it as vulnerable. Even earlier protection tools such as ProPolice and

StackShield failed to detect [46]. It is also difficult to exploit since ones need to know the exact length of character, address of function call, and address of environment variable

A sample vulnerable code contains return-into-libc vulnerability.

```
int main(char **av)
{
    char a[20];
    if ( strlen(av[1]) < 20 ) //verified the length
       strcpy(a, av[1]);      //nothing happen
    printf ("%s", a);  //possible have vulnerability
    return 0;
}
```

Based on the code above, it is possible to supply a string long enough to fill up the allocated memory space together with function call to replace the defined return address. The result of exploiting it is extremely dangerous [47]. To recognize the vulnerability, security analysts need to understand possible input values and estimate memory location. It is similar to Unsafe Function and Array Out-of-bound class but differ in terms of behavior and memory process. Memory in the former two classes will overflow and overwrite the return address, resulting in unintended behavior. In contrast, Return-into-lib will replace return address with a function call to another program e.g. *system()* and *WinExec()* [48], [49], [50]. To prevent it from appearing or being exploited, the contents of the input must be validated apart from the length.

3.5 Memory Function

Even though it has been in the security radar as early as 2002 [52], [53], [54], [55], [56], [57], [58], [59], it is not been properly classified. This type of vulnerability has gain notoriety as one of the preferred vulnerability for exploitation due to current programming trend as more programs are developed to utilize dynamic memory for better performance and scalability.

Double call on *free()* function, improper use of *malloc()*, *calloc()*, and *realloc()* functions, uninitialized memory, and unused allocated memory are few examples of memory functions vulnerabilities. A simple memory function vulnerability is shown below.

A fragment of C code with *free()* function vulnerability.

```
char* ptr = (char*) malloc (DEFINED_SIZE);
...
free(ptr);  //first call to free ptr
free(ptr);  //vulnerable due to free the freed ptr
```

As shown above, the second call to free the same variable will cause unknown behavior. This can be used for exploitation and its severity is equivalent to the first three types [61], [60].

Due to its safe nature and programming complexity, it is difficult to assess its vulnerability potential unless an in-depth semantics view of program is used. From coding perspective, this type of vulnerability can be prevented by validating the memory before usage, initializing memory with default value depending on variable type, and removing unused memory.

3.6 Function Pointer / Pointer Aliasing

Function pointer or pointer aliasing is a variable storing address of a function or as reference to another variable. It can later be called by the given pointer name which assists developer's flexibility and ease of programming. It becomes vulnerable when the reference has been nullified or overwritten to point to a different location or function [52], [62], [63], [64].

It is difficult to detect by manual code review unless it is done by highly experience security analysts. However, using automatic tool requires the tool to comprehend semantically the code [65]. Below is an example.

An example of pointer aliasing vulnerability.

```
char s[20], *ptr, s2[20];
ptr = s;              //vulnerable line of code
strncpy(s2, ptr, 20); // vulnerability realized
```

As shown above, the pointer *ptr* is referring to a null value since the variable *s* is yet to be initialized. The subsequent line of code realizes the vulnerability, although the function used is a safe function. The only way to stop this type of vulnerability from continuing to occur is by enforcing validation on pointer variable before being used throughout the program.

3.7 Variable Type Conversion

Improper conversion of a variable can create vulnerability and exploitation [67], [68], [69]. Although there are considerable numbers of advisories reporting this vulnerability [67], [70], it was never mentioned in any earlier taxonomy. It may be due to infrequent occurrence and minimal severity. It was also considered as a member of integer overflow vulnerability which is arguable since conversion errors do happen on other data format. A true example of this vulnerability is shown below.

Fraction of Bash version 1.14.6 contains Variable Type Conversion vulnerability [73].

```
static int yy_string_get() {
  register char *string;
  register int c;

  string = bash_input.location.string;
  c = EOF;

  ......
  if (string && *string) {
    c = *string++;        //vulnerable line
    bash_input.location.string = string;
  }
  return (c);
}
```

This vulnerability is proven to be exploitable [67], [68], [69], [71], and [72] and ignoring it is reasonably risky. To avoid conversion error vulnerability, it is strongly suggested to validate all variable involves in conversion, as well as avoid unnecessary conversion, or use the same data type.

3.8 Pointer Scaling / Mixing

This vulnerability may arise during an arithmetic operation of a pointer [74], [75]. Semantically, it is different to pointer aliasing in terms of coding. It seldom happens but the impact of exploiting it is comparable to other type of overflow vulnerability.

In a pointer scaling or mixing process, the size of object pointed will determine the size of the value to be added [75]. If one failed to understand this, he or she may wrongly calculate and assign the wrong size of object to accept the result, and therefore runs the risk of having overflow vulnerability.

Sample of code contains Pointer Scaling / Mixing vulnerability [76].

```
int *p = x;
char * second_char = (char *)(p + 1);
```

As shown in the above code, subsequent read or write to pointer *second_char* will cause overflow or unknown behavior due to addition of value 1 to current address location of variable *x*.

To avoid this vulnerability from occurring, ones must recognize and be able to correctly determine the accurate size of recipient variable and actual location in memory.

3.9 Uninitialized Variable

Uninitialized variable is variable declared without value assigned to it. Nonetheless, computer will allocate memory and assign unknown values, which later if being used will cause computer system to perform undesired behavior [77], [78], [79]. It can also be exploited by attackers thus allowed the system to be compromised [80].

A fraction of C code contains Uninitialized Variable vulnerability [80].

```
....
void take_ptr(int * bptr){
    print (“%lx″, *bptr);
}
```

```
int main(int argc, char **argv){
    int b;
    take_ptr(&b);
    print (“%lx”, b);
}
```

Variable *b* in the sample above was not initialized and then being used twice without value assigned to it. By default, a location has been set in memory and the next line after declaration will force either computer system to behave abnormal or be vulnerable for exploitation. Normal compiler, code review, or even most static analysis will not mark this as vulnerable. There is also possibility of triggering false alarm if there is value assign to it before use.

The easiest way to overcome this vulnerability is to initialize all variable with acceptable default value such as zero for numerical type of variable and empty string or blank space for character or string. It must not be left uninitialized or contain null value before used. Another way to avoid this vulnerability which might impact performance is to validate variable before usage.

3.10 Null Termination

Although it seem likely to ensue and can easily be avoided, it still appear in few vulnerability databases [82] and [83]. The consequence of having this vulnerability is equally dangerous as other type of vulnerabilities [81].

Null termination vulnerability is defined as improper string termination, array that does not contain null character or equivalent terminator, or no null byte termination with possible impact of causing overflows [52], [54], [81]. A sample of this vulnerability is shown in code below.

Fraction of C code contains Null Termination vulnerability [81].

```
#define MAXLEN 1024
...
char *pathbuf[MAXLEN];
...
read(cfgfile,inputbuf,MAXLEN);
strcpy(pathbuf, inputbuf);
```

The above code which seems safe as the *read()* function has limited the size of input to the same size of destination buffer on the last line. However, if the input did not have null termination, due to behavior of *strcpy()*, it will continue to read it until it find a null character. This makes it possible to trigger an overflow on the next reading of memory. Even if it was replaced with *strncpy()*, which is considered as safe, the behavior is still unpredictable, thus making it a unique vulnerability on its own.

To overcome the vulnerability, one should validate the input before use, or restrict the length of input to have less than one from the actual defined size.

4 Summary of Taxonomy on C Code Overflow Vulnerabilities Attack

Table 2 summarizes our proposed taxonomy, consisting of ten categories of C code overflow vulnerability attacks, their severity, likelihood to appear and mode of exploitation. The occurrence and severity of listed vulnerability type is based on our thorough evaluation on various advisories and reports by [7], [8], [10], and [11].

Table 2. Summary of Taxonomy on C Code Overflow Vulnerability Attack

Overflow Type	Mode of Exploit	Code Appearance	Severity	Occurrence
Unsafe Function	Supplying malicious input long enough to overwrite memory location	No validation on input before being used in unsafe function or restricting unsafe function	Critical	High
Array Out-of-Bound	Supplying input or forcing access on array beyond defined index either below minimum or above minimum index.	No validation on index of array before being used.	Critical	High
Integer Range/Overflow	Supplying input used in arithmetic operation forcing the result to overwrite memory defined or exploiting miscalculation of arithmetic operation	Improper estimation on result of arithmetic calculation	Critical	High
Return-into-libc	Overwriting return address with address of library function	Uncheck argument passing in a function call	Critical	Low
Memory Function	Exploiting misuse of memory function (i.e. double call to *free()*)	Never use allocated memory, double free of same memory or calling freed memory.	Critical	Medium
Function Pointer / Pointer Aliasing	Overwriting the function pointer to point address that contains malicious code or function	Use of pointer without validating the pointer first	Medium	Medium
Variable Type Conversion	Exploiting vulnerabilities exist during conversion of different variable type	Miscalculation of variable size involves in conversion	Medium	Low
Pointer Scaling / Pointer Mixing	Exploiting vulnerabilities trigger during arithmetic operation of a pointer	Miscalculation of pointer size in scaling or mixing process	Medium	Low
Uninitialized Variable	Exploiting vulnerabilities when uninitialized variable being used in the program	A variable being used before initialization	Medium	Low
Null Termination	Supplying non-terminated input	No null termination validation on input	Medium	Low

5 Conclusion

We have discussed various classifications of software overflow vulnerabilities, and presented the strengths and weaknesses of previous taxonomies in general, and overflow and C vulnerabilities in particular. We noted at present there is no taxonomy specifically addressing overflow vulnerabilities from C source code perspective. Therefore, we construct taxonomy for C overflow vulnerabilities attack. In producing this taxonomy, we focus on how the overflow vulnerability appears in C code and the criteria used for a code to be considered as vulnerable. We demonstrated the application of our taxonomy in identifying types of C overflow vulnerabilities by providing a few sample vulnerable code segments. The taxonomy can be a valuable reference for developers and security analysts to identify potential security C loopholes so as to reduce or prevent exploitations altogether.

6 Future Work

We look forward to validate and verify our taxonomy with standard vulnerability databases and implement it to evaluate the effectiveness of the security vulnerability program analysis tools.

References

1. Aleph One: Smashing the Stack for Fun and Profit. Phrack Magazine 7(49) (1996)
2. Zitser, M.: Securing Software: An Evaluation of Static Source Code Analyzers. M. Sc. Thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2003)
3. Kratkiewicz, K.: Evaluating Static Analysis Tools for Detecting Buffer Overflows in C Code. M. Sc. Thesis. Harvard University (2005)
4. Zhivich, M.A.: Detecting Buffer Overflows Using Testcase Synthesis and Code Instrumentation. M. Sc. Thesis. Massachusetts Institute of Technology (2005)
5. Akritidis, P., Cadar, C., Raiciu, C., Costa, M., Castro, M.: Preventing Memory Error Exploits with WIT. In: IEEE Symposium on Security and Privacy, pp. 263–277. IEEE Computer Society, Washington, DC, USA (2008)
6. Common Vulnerability and Exposures, `http://cve.mitre.org/`
7. Microsoft Security Advisories, `http://www.microsoft.com/technet/security/advisory`
8. IBM X-Force Threat Reports, `https://www-935.ibm.com/services/us/iss/xforce/trendreports/`
9. 2010 CWE/SANS Top 25 Most Dangerous Software Errors, `http://cwe.mitre.org/top25/`
10. Buffer Overflow on Common Vulnerability and Exposures, `http://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=Buffer+Overflow`
11. Microsoft Security Advisories Archive, `http://www.microsoft.com/technet/security/advisory/archive.mspx`

12. Chess, B., McGraw, G.: Static Analysis for Security. J. IEEE Security and Privacy. 2(6), 76–79 (2004)
13. Foster, J.S., Hicks, M.W., Pugh, W.: Improving software quality with static analysis. In: 7th ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering, pp. 83–84. ACM, New York (2007)
14. Emanuelsson, P., Nilsson, U.: A Comparative Study of Industrial Static Analysis Tools. J. Electronic Notes in Theoretical Computer Science (ENTCS) 217, 5–21 (2008)
15. Howard, M., LeBlanc, D., Viega, J.: 24 Deadly Sins of Software Security: Programming Flaws and How to Fix Them. McGraw Hill, United States of America (2009)
16. Viega, J., McGraw, G.: Building Secure Software: How to Avoid Security Problems the Right Way. Addison-Wesley Professional, United States of America (2001)
17. Seacord, R.C.: Secure Coding in C and C++. Addison-Wesley Professional, United States of America (2005)
18. Krsul, I.V.: Software Vulnerability Analysis. Phd. Thesis. Purdue University (1998)
19. Lough, D.L.: A Taxonomy of Computer Attacks with Applications to Wireless Networks. Phd. Thesis. Virginia Polytechnic Institute and State University (2001)
20. Aslam, T.: A Taxonomy of Security Faults in the UNIX Operating System. M. Sc. Thesis. Department of Computer Science, Purdue University (1995)
21. Alhazmi, O.H., Woo, S.W., Malaiya, Y.K.: Security Vulnerability Categories in Major Software Systems. In: 3rd IASTED International Conference on Communication, Network, and Information Security (CNIS) ACTA Press, Cambridge (2006)
22. Pothamsetty, V., Akyol, B.: A Vulnerability Taxonomy for Network Protocols: Corresponding Engineering Best Practice Countermeasures. In: IASTED International Conference on Communications, Internet, and Information Technology (CIIT). ACTA Press, US Virgin Islands (2004)
23. Bazaz, A., Arthur, J.D.: Towards A Taxonomy of Vulnerabilities. In: 40th International Conference on System Sciences, Hawaii (2007)
24. Gegick, M., Williams, L.: Matching Attack Patterns to Security Vulnerabilities in Software-Intensive System Designs. In: Workshop on Software Engineering for Secure Systems – Building Trustworthy Applications. ACM, New York (2005)
25. Howard, J.D., Longstaff, T.A.: A Common Language for Computer Security Incidents. Sandia Report (SAND98-8667). Sandia National Laboratories, California (1998)
26. Tsipenyuk, K., Chess, B., McGraw, G.: Seven Pernicious Kingdoms: A Taxonomy of Software Security Errors. IEEE Security and Privacy 3(6), 81–84 (2005)
27. Hansman, S., Hunt, R.: A taxonomy of network and computer attacks. J. Computer and Security. 24(1), 31–43 (2005)
28. Hansmann, S.: A Taxonomy of Network and Computer Attacks Methodologies. Technical Report. Department of Computer Science and Software Engineering, University of Canterbury, New Zealand (2003)
29. Killourhy, K.S., Maxion, R.A., Tan, K.M.C.: A Defense-Centric Taxonomy Based on Attack Manifestations. In: International Conference on Dependable Systems and Networks, pp. 91–100. IEEE Press, Los Alamitos (2004)
30. Igure, V., Williams, R.: Taxonomies of Attacks and Vulnerabilities in Computer Systems. J. IEEE Communications Surveys and Tutorials. 10(1), 6–19 (2008)
31. Shahriar, H., Zulkernine, M.: Taxonomy and Classification of Automatic Monitoring of Program Security Vulnerability Exploitations. J. Systems and Software 84, 250–269 (2011)
32. Sotirov, A.I.: Automatic Vulnerability Detection Using Static Source Code Analysis. M. Sc. Thesis. University of Alabama (2005)

33. Moore, H.D.: Exploiting Vulnerabilities. In: Secure Application Development (SECAPPDEV). Secappdev.org (2007)
34. Metasploit Penetration Testing Framework, http://www.metasploit.com/framework/modules/
35. Symantec Threat Explorer, http://www.symantec.com/business/security_response/threatexplorer/vulnerabilities.jsp
36. Wagner, D.: Static Analysis and Computer Security: New Techniques for Software Assurance. Phd. Thesis. University of California, Berkeley (2000)
37. Security Development Lifecycle (SDL) Banned Function Calls, http://msdn.microsoft.com/en-us/library/bb288454.aspx
38. Stanford University: Pintos Project, http://www.stanford.edu/class/cs140/projects/pintos/pintos.html#SEC_Top
39. Secunia Advisories, http://secunia.com/advisories/
40. Engler, D.: How to find lots of bugs in real code with system-specific static analysis, http://www.stanford.edu/class/cs343/mc-cs343.pdf
41. Ashcraft, K., Engler, D.: Using Programmer-written Compiler Extensions to Catch Security Holes. In: IEEE Symposium on Security and Privacy, pp. 143–159 (2002)
42. Red Hat Bugzilla: Bug 546621, https://bugzilla.redhat.com/show_bug.cgi?id=546621
43. National Vulnerability Database: Vulnerability Summary for CVE-2010-4409, http://web.nvd.nist.gov/view/vuln/detail?vulnId=CVE-2010-4409
44. Integer Overflow, http://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=Integer+Overflow
45. CWE-190: Integer Overflow or Wraparound, http://cwe.mitre.org/data/definitions/190.html
46. Richarte, G.: Multiple Vulnerabilities in Stack Smashing Protection Technologies. Security Advisory, Core Labs (2002)
47. Stack Overflow, http://www.owasp.org/index.php/Stack_overflow
48. Lhee, K., Chapin, S.J.: Type-Assisted Dynamic Buffer Overflow Detection. In: 11th USENIX Security Symposium. USENIX Association, CA (2002)
49. Nelißen, J.: Buffer Overflows for Dummies. SANS InfoSec Reading Room - Threats/Vulnerabilities. SANS Institute (2003)
50. Nergal: The Advanced Return-into-lib(c) Exploits. Phrack Magazine 11(58) (2001)
51. Using Environment for Returning Into Lib C, http://www.securiteam.com/securityreviews/5HP020A6MG.html
52. Grenier, L.A.: Practical Code Auditing. Metasploit Framework (2002)
53. Akritidis, P., Cadar, C., Raiciu, C., Costa, M., Castro, M.: Preventing Memory Error Exploits with WIT. In: IEEE Symposium on Security and Privacy, pp. 263–277 (2008)
54. Tevis, J.J., Hamilton, J.A.: Methods for the Prevention, Detection and Removal of Software Security Vulnerabilities. In: 42nd Annual Southeast Regional Conference, pp. 197–202 (2004)
55. SecurityFocus, http://www.securityfocus.com/archive/1/515362
56. Microsoft Security Bulletin MS03-029, http://www.microsoft.com/technet/security/bulletin/ms03-029.mspx

57. iDefense Labs Public Advisory (December 6, 2007), http://labs.idefense.com/intelligence/vulnerabilities/display.php?id=542
58. CVE-2005-3828, http://www.cvedetails.com/cve/CVE-2005-3848/
59. Testing for Heap Overflow, http://www.owasp.org/index.php/Testing_for_Heap_Overflow
60. Double Free, http://www.owasp.org/index.php/Double_Free
61. CWE-415: Double Free, http://cwe.mitre.org/data/definitions/415.html
62. Kolmonen, L.: Securing Network Software using Static Analysis. In: Seminar on Network Security, Helsinki University of Technology (2007)
63. Nagy, C., Mancoridis, S.: Static Security Analysis Based on Input-related Software Faults. In: European Conference on Software Maintenance and Reengineering, pp. 37–46. IEEE Computer Society, Los Alamitos (2009)
64. Durden, T.: Automated Vulnerability Auditing in Machine Code. Phrack Magazine (64) (2007)
65. Michael, C., Lavenhar, S.R.: Source Code Analysis Tools – Overview. Homeland Security (2006)
66. Wagner, D., Foster, J.S., Brewer, E.A., Aiken, A.: A First Step Towards Automated Detection of Buffer Overrun Vulnerabilities. In: Network and Distributed System Security (2000)
67. C Language Issues for Application Security, http://www.informit.com/articles/article.aspx?p=686170&seqNum=6
68. Pozza, D., Sisto, R.: A Lightweight Security Analyzer inside GCC. In: 3rd International Conference on Availability, Reliability and Security, Barcelona, pp. 851–858 (2008)
69. Morin, J.: Type Conversion Errors. In: Black Hat, USA (2007)
70. FFmpeg Type Conversion Vulnerability, http://securityreason.com/securityalert/5033
71. CWE-704: Incorrect Type Conversion or Cast, http://cwe.mitre.org/data/definitions/704.html
72. CWE-195: Signed to Unsigned Conversion Error, http://cwe.mitre.org/data/definitions/195.html
73. STR34-C. Cast characters to unsigned char before converting to larger integer sizes, https://www.securecoding.cert.org/confluence/display/seccode/STR34-C.+Cast+characters+to+unsigned+char+before+converting+to+larger+integer+sizes
74. Black, P.E., Kass, M., Kog, M.: Source Code Security Analysis Tool Functional Specification Version 1.0. In: NIST Special Publication 500-268 (2007)
75. Seacord, R.C.: The CERT C Secure Coding Standard. Addison-Wesley Professional, Reading (2008)
76. Unintentional Pointer Scaling, http://www.owasp.org/index.php/Unintentional_pointer_scaling
77. Uninitialized variable, http://en.wikipedia.org/wiki/Uninitialized_variable
78. Eight C++ programming mistakes the compiler won’t catch, http://www.learncpp.com/cpp-programming/eight-c-programming-mistakes-the-compiler-wont-catch/

79. Uninitialized Variable, http://www.owasp.org/index.php/Uninitialized_Variable
80. Flake, H.: Attacks on Uninitialized Local Variables. Black Hat Federal (2006)
81. CWE-170: Improper Null Termination, http://cwe.mitre.org/data/definitions/170.html
82. Microsoft Security Bulletin MS09-056, http://www.microsoft.com/technet/security/bulletin/ms09-056.mspx
83. CVE-2007-0042, http://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2007-0042
84. SolarDesigner: Getting around non-executable stack (and fix). Bugtraq Mailing List, http://www.securityfocus.com/archive/1/7480

A Secure Storage Model to Preserve Evidence in Network Forensics

Mohd Izham Ibrahim and Aman Jantan

Security Research Group, School of Computer Sciences,
Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia
mii10_com053@student.usm.my, aman@cs.usm.my

Abstract. The main challenge in Network Forensics, especially during the Trial session, is to protect the evidences and preserve the contents from malicious attempts to modify and tamper it. Any potential evidences that are not accurate, complete, reliable and verifiable will certainly affect the decision among the jury and judges. In this paper, we classify the potential evidences that will be stored in the network storage based on their contents, characteristics and functions. We also propose a Secure Storage Model, which implements components that preserve evidences using Cryptographic Hashing and Logging Report. As a result, we present the flow of our storage mechanisms and show the importance of hashing for forensics work to secure collected network evidences.

Keywords: network forensics, secure storage, network digital evidence, cryptographic hash, preserving evidence.

1 Introduction

A digital forensics investigation is a process that investigates the crimes conducted in digital environments with the interest of legal system [1], [2], [3]. The sub-discipline of this procedure is called Network Forensics (NF) because works in network-based environments monitoring network traffic to detect intrusion, collecting legal evidences and detecting misuse of network resources [4]. However, network forensics seems to face serious issues resolving to the collected evidence from network traffic as the evidences is vulnerable and unpredictable [5], [6], [7]. Therefore, once the evidences were collected, any sort of glitch detected will turn it inadmissible in Courts. Whilst not all of the evidences will be useful thus it is up to the forensics investigators to filter and present the evidence in Courts. Yet, the chain-of-custody process and the integrity of the evidences would still be the main important thing in any Trial session.

Nowadays, to support the Trial sessions for digital crimes investigation, there are several methodologies [8], [9], [10], [11], [12] that assist the investigation process. The common practices of this methodology are not consistent and different approach was based on the experience of several people such as law enforcement authorities, the hackers and also system administrator [13]. However, most of the methodology would share the common steps as shown in Fig. 1. These include Collection, Analysis and Presentation.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 391–402, 2011.

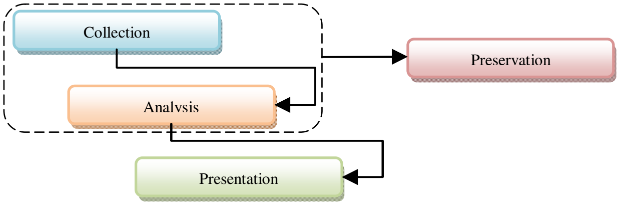


Fig. 1. Common Digital Forensics Methodology

The common methodology indicates that, preservation is something that should be taken care of from the first phase of investigation, which is Collection phase and also during the Analysis phase. This is to protect the integrity of the evidence that was collected and to ensure that the Chain-of-custody of the evidence is not broken while the investigators examine the evidence. A very common question asked while Trial session would be, Was the data altered? [14]. Thus, to make sure that the evidence was not altered during the process, researchers have come out with various types of methods to prevent alteration such as documenting each procedure taken, creating a secure architecture [15], calculating the hash value of the evidence, creating a copy of evidences for analysis purpose, encrypting the contents and comparing the hash value for later recognition that the evidence was not tampered [4], [16]. However, the preservation phase is an iterative process that goes along in each phase that requires the evidence to be preserved in every phase along the period of investigations.

In network-based environments, the concepts of preserving the evidence is much more complicated [17] as once the evidence was collected from the network traffic, it is opened to tampering and modification either stored in local storage or any network storage [18]. It is possible that, the mechanism to store the potential evidence and also the storage container can be argued in courts, if the chain-of-custody is deniable and proved to be wrong.

In order to further understand how methods defined in network forensic to collect and preserving the evidence, we do briefly study on past work in Section 2. We found out there are problems related to preserving evidence. Based on our findings of these problems, we proposed our own model to secure the storage in Section 3 with our proposed storing mechanism and applying hashing functions to protect the integrity of our stored evidence. Towards the end of this paper, we conclude our work and define our expectation for the future work.

2 Related Works

In general, network traffic produce lots of information that can turn out to be evidences if there is a report about crimes happening. [1], [13], [15], [18], [19], [20], [21] describe the basic problem when handling with evidences in network-based environments is to acquire and process massive amounts of digital evidence while preserving the integrity of each evidences for extensive time. David, M. [18] investigate network-based architectures for storing and handling large amounts of evidence by proposing Digital Evidence Storage Lockers (DESLs), which implements Network Attached Storage (NAS) over Storage Area Network (SAN).

DESLs works as a centralized repository where evidence is stored and maintained in a remote locations which allows secure access. However, Mahalingam, P. [21] address about the security concerns of protecting the transmission of SAN, among the host and storage device since SAN communicates the block level data transfer. According to Mahalingam, 18% are not willing to implement SAN due to its security problems and those who already have, 56% expressed security concerns. In most cases, the SAN faces three types of threats, 1) malicious outsider's threats, 2) malicious insider's threats and 3) non-malicious insider threats [21], [22]. The threats to the evidences is very dangerous as it could lead to unauthorized reading of data, unauthorized alteration of data and unauthorized destruction of data [23]. However, [24] mentioned about the levels of protection against these threats starting from database system, operating system, network, physical and human. Due to each level protection, evidence can be found once the protection was breach by attackers, synonym to traditional crimes as the criminals often leave something behind.

Evidences that was left behind, is either with or without the criminals acknowledgment. According to Sommer, P. [25], the criminals who often leave something behind is different based on the location of crimes either host based or network based. In host-based location, the use of Disk Forensics to extract the evidences from the hard-disk by first make an exact copy of the hard-disk and then analyze the content for any hidden or deleted contents suit with crime case reported. However, evidences in network-based location need to be captured from the network and they will match it against individual computers to justify for the crimes conducted [25], [26].

There is a lot of potentials evidence that can be retrieved from each level of protection. According to Kozushko, H. [27], the evidence found can be organize in three classifications which are 1) by content 2) by functions 3) by characteristics. Evidences that was classify by contents will contain the information about the owner of the evidence and moreover, digital evidences that come from swap files and slack space, often contains fragments of digital data that can be reorganized to create a much more useful information. Other than that, evidences can be classify by functions whereby, the functions will indicates what type of threats it can harm the systems, such as malware infection over the network. The last classification define by Kozushko, is characteristics, lots of information over evidences can be gather by knowing the characteristic of certain evidence, such as the file names, message digests and data stamp [27]. However, in network-based the source of evidence can exist in many different formats depends on the layer in OSI Model [28]. The source can be from –packet filter logs, proxy logs, sniffers, DNS cache, e-mail messages, browser histories, cookie and cache [28], [29].

Previous researchers discuss briefly about requirements of evidence that are admissible in Courts. According to Accorsi, R. [29], "admissibility is a legal concept that prescribes requirements to judge the acceptance of any kind of evidence". Thus, to place evidences in the Storage, the mechanism need to be 1) *entry accountability* which include information with regard to the subject, 2) *entry integrity* as the list of stored evidence were not modified (*accuracy*) or the list was not deleted (*completeness*). The last requirement defined by Accorsi, R. was *entry confidentiality* as the entries are not stored in clear-text making sure that the contents were not readable by others. Other than that, other researchers [30], [31], [32], proposed that, the evidence also need to be authentic, complete, reliable and believable. However, all the evidences that were presented by an expert in the area would always get the evidences admissible in Courts unless the chain-of-custody can be prove broken during the investigations.

Chain-of-custody was preserved from the time the evidence was collected until the time it was presented. According to Scarlet, S.D [33], the most important thing is to protect the chain-of-custody of the best evidence and locked with key. Hosmer, C. [14] address concerns of proving the integrity of the original evidence by checking the hash value of the best evidence's copy which is the first copy of the evidence and compare with the one presented in Courts. If the two values were not match, the integrity or reliability of the copied evidence is questionable.

A cryptographic hash function is a common practice in any investigation to proof the integrity of the digital evidence. According to Roussev, V. [34], "hashing is a primary, yet underappreciated, tool in digital forensic investigations". Sometimes, hashing might as well get neglected and was analyzed right away after it was transferred from the crime scene. There are possibilities of the evidence to be tampered and no records were found to support in front of Courts. In crimes investigation, the process of encrypt and decrypt was used to hide the real data from non-intent viewers as it protects the privacy and confidentiality of the evidence, yet the integrity was not preserved. So hashing will do all the work of keeping the integrity of best evidence's copy with the original evidence. Hash is a "signature" for a stream of data that represents the contents [35]. The size of contents might be large and yet produce small output of hash value. However, cryptographic hashing is different from encryption, whereby hashing use one-way operation instead of two-way operation for encryption. Other than that, the two way operation of encrypt and decrypt would produce the same amount of size as illustrated in Fig. 2.

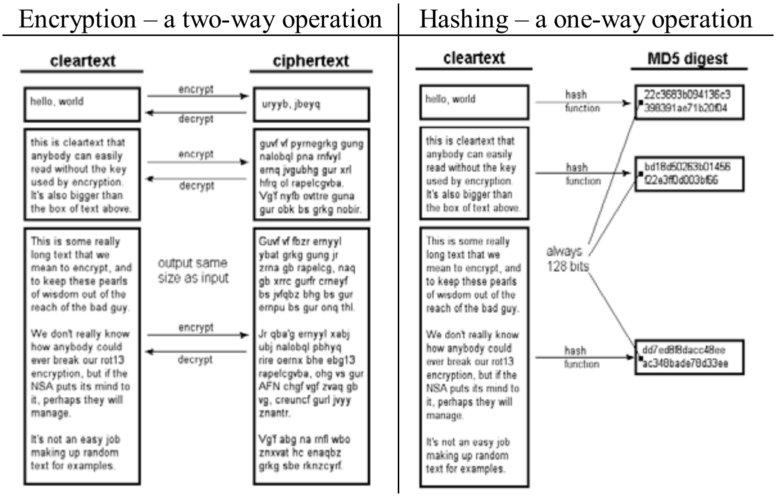

Encryption – a two-way operation
Hashing – a one-way operation
cleartext
ciphertext
hello, world
encrypt
decrypt
uryyb, jbeyq
this is cleartext that anybody can easily read without the key used by encryption. It's also bigger than the box of text above.
encrypt
decrypt
guvf vf pyrnegrkg gung nalobql pna rnfvyl ernq jvgubhg gur xrl hfrq ol rapelcgvba. Vg'f nyfb ovttre guna gur obk bs grkg nobir.
This is some really long text that we mean to encrypt, and to keep these pearls of wisdom out of the reach of the bad guy.
We don't really know how anybody could ever break our rot13 encryption, but if the NSA puts its mind to it, perhaps they will manage.
It's not an easy job making up random text for examples.
output same size as input
encrypt
decrypt
Guvf vf fbzr ernyyl ybat grkg gung jr zrna gb rapelcg, naq gb xrrc gurfr crneyf bs jvfqbz bhg bs gur ernpu bs gur onq thl.
Jr qba'g ernyyl xabj ubj nalobql pbhyq rire oernx bhe ebg13 rapelcgvba, ohg vs gur AFN chgf vgf zvaq gb vg, creuncf gurl jvyy znantr.
Vg'f abg na rnfl wbo znxvat hc enaqbz grkg sbe rknzcyrf.
cleartext
MD5 digest
hello, world
hash function
22c3683b094136c3 398391ae71b20f04
this is cleartext that anybody can easily read without the key used by encryption. It's also bigger than the box of text above.
hash function
bd18d50263b01456 f22e3ff0d003bf66
This is some really long text that we mean to encrypt, and to keep these pearls of wisdom out of the reach of the bad guy.
We don't really know how anybody could ever break our rot13 encryption, but if the NSA puts its mind to it, perhaps they will manage.
It's not an easy job making up random text for examples.
always 128 bits
hash function
dd7ed8f8dacc48ee ac348bade78d33ee

Fig. 2. The different of two-way operation and one-way operation with MD5 digest [35]

The most commonly used cryptographic hash functions were MD5 and SHA-1. [38]. However, the algorithm for the hash function is not limited to this two only, based on Fig. 3, there are few other functions developed from 1989 and their popularity was base on the weakness as they can be attack and manipulate. According to Roussev, V. [34], some hash function is collision resistant meaning that it's very hard to find similar hash output from this function such as; MD5, RIPEMD-160, SHA-1, SHA-256, and SHA-512 as they produce large size of bit [34]. Large size of bit means that, the length of the digest output is very long and chances of having similar output would be rare. However, the risk could still be there and to minimize it, the use of several functions together at a time would be recommended.

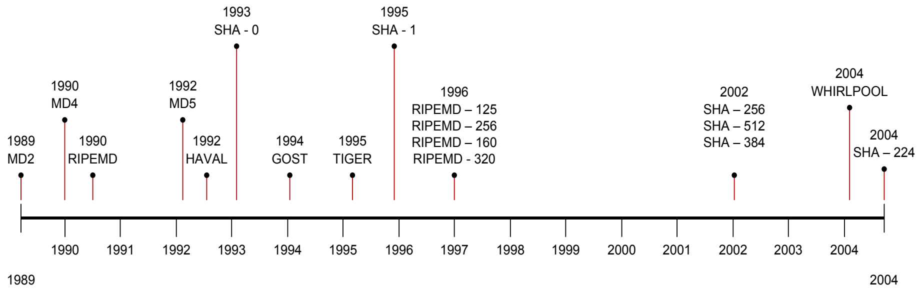


Fig. 3. Timeline for hash functions

Logging report on the other hand, is also an important factor in any investigation because it will show the details of the evidence from the time it was collected and until it was presented. According to Cosic, J. [37], it is important to track the evidence material at all time and access of the evidence must be controlled and audited. Cosic, J. also proposed Digital Evidence Management Framework (DEMF) using the "Five Ws (and one H)" – Who, What, When, Where, Why, and How. which can be presented as a function [37]

```
DEMF = f
      {
              fingerprint_file, //what
              biometric characteristic, //who
              time_stamp, //when
              gps_location, //where
      };.
```

Time stamp is a good indicator to prove the integrity of the evidence. According to Hosmer, C. [14], a secure and auditable time can eliminates the potential for fraud and unintended errors. Most of this technique have been covered in commercial tools such as AccessData's FTK, Guidance Software's EnCase and open source tools such as SleuthKit, yet, none had provided secure mechanism for the storage of the evidence.

3 Secure Storage Model for Network Evidence

In this section, we discuss about the proposed model to secure the storage for forensics work. We proposed the general architecture of our model and elaborate on the storing mechanism that we use. We also explain about key components of this model with details on the cryptographic hash engine and logging report that we implement in the model to protect the integrity of the evidence until it was called for presentation.

3.1 General Architecture

As mentioned earlier, the methodology for Network Forensics can be divided into three steps; Collection, Analysis and Presentation. We proposed a secure model by dumping the evidence into these steps to hold the evidence. The model is shown in Fig. 4.

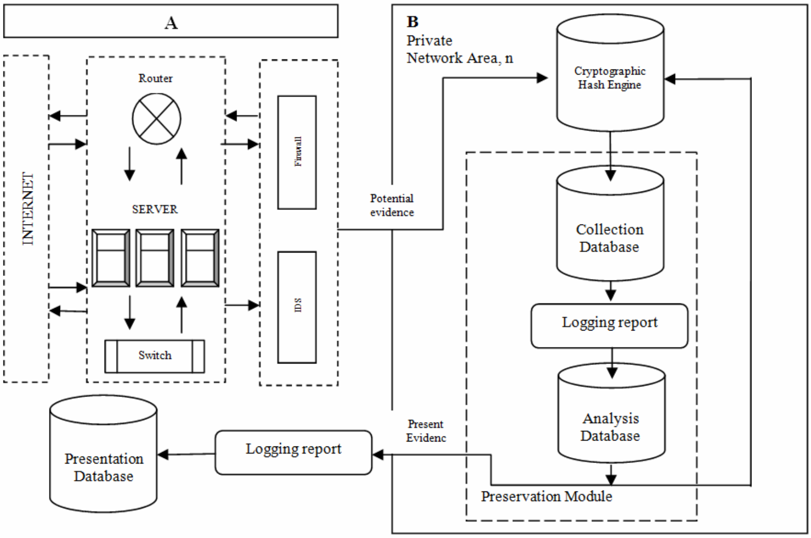


Fig. 4. The proposed Secure Storage Model for Network Forensics work

The proposed model was separated into two sector; A and B, where sector A is the general location in the network environment that contains potential evidences. However, sector A is not briefly discussed in this paper because the focus of this paper is about the mechanism to store the evidence collected from sector A. Yet, sector A will provide lots of evidence while Sector B is important to classify the evidence and store it for analysis purpose by the investigators. However, we do realize the importance of having Firewall and Intrusion Detection System (IDS) from sector A in our model because, the capability of these tools to detect and filter any attacks or intrusion. Sector B in this model is one of many other Private Network Area based on the Case handled by the investigator. We purposely create new Network Area for each Case to make it more individualize while handling the evidence.

The Private Network Area (PNA) contains Cryptographic Hashing (CH) engine, logging report (LR) and databases of the evidences from the time it was collected until it was analyze and presented. LR act as a tool to monitor and records the investigators involve collecting the evidence, and provide additional information such as hash value generated from the CH engine, time, location and places that related to the evidences stored. It also operates as a reporting tool for any attempts and action to access the evidence while the Case is still ongoing.

3.2 Key Components

Storing Mechanism and Logging Report (LR). We illustrated the mechanism of storing the evidence by using a flowchart in Fig. 5. When a Case was issued, the investigators will seek out sources for potential evidences. The source came from

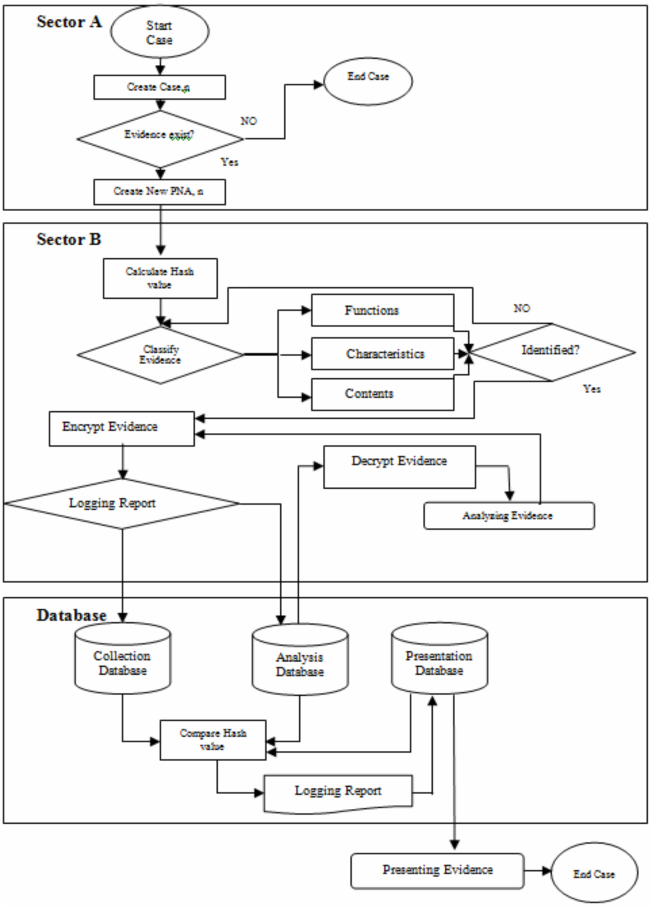


Fig. 5. Flowchart of Evidence Storing Mechanism

network traffic for example firewall logs, intrusion detection system logs, network communications link records or maybe information on network device such as routers, dial-up servers, switches and etc [36]. If there is no evidence exists, the Case will be drop and the investigations process will be stop. However, if the evidence exists, a new PNA will be created and the hash value of the evidence will be process. We discuss more about the CH used for this model in next section.

After that, our model classifies and separate the evidences based on the classification of the evidence. The reason to classify the evidence is to make sure important evidence based on the functions, characteristic and contents is identified and prepared a suitable location to store it. For example, a file that contains potential logic bomb characteristic that can trigger and assume can delete other files should be identified earlier and kept out of other file's reach. The same things with evidence that contains malicious contents and have the same characteristic as a malware should be identified.

The next part involves encrypting the evidence to hide the contents and to protect the confidentiality of the evidence collected. This is important as the evidence collected might contain sensitive data and not open to public even in the event that a database is successfully attacked or stolen [36]. By encrypting the evidence, potential exposure of this private data can be reduce and only authorized law officer or investigators can access the evidence. Each of the process and time will be logged in the LR and then the evidences will be copied inside different database. For analyzing purpose, the investigators will need to decrypt the evidence and retrieve the evidence for analyze purpose from the Analysis Database, as this is reserved for the Investigators to retrieve needed information as the other Collection Database already store the copied evidence.

The last part of this mechanism ends when the Investigators found out the needed information and compare the hash value of the evidence. Then all of the work will be recorded inside LR for reporting and save inside Presentation Database. During Trial session, the needed information will be requested from the Presentation Database and after Presenting the evidence the mechanism will end its process.

Cryptographic hashing Mechanism. In general, cryptographic hashing (CH) is commonly used to protect the integrity of file from any attacks to tamper and alter the file. From 1990's, there a lot of CH being developed and the most well-known would be Message Digest Algorithm 5 (MD5) [9]. However, because of flaw issues and the weakness of MD5 [35], [37], [38], [39], [41] we proceed to use Whirlpool [40] to digest the evidence and retrieve the hash value. Whirlpool was recommended by New European Schemes for Signatures, Integrity and Encryption (NESSIE) and also used by International Organization for Standardization (ISO) and the International Electrotechnical Commision (IEC) [41]. The credibility of these institutions would be beneficial in Courts to defend the evidence presented.

Another factor to use Whirpool is because it produces 512 bits output size. It is important to have large output to reduce collision among outputs when presented. Using an sample of a pcap file, we calculate the hash value of this file and illustrated the use of Whirlpool in protecting the integrity of evidence. Using Febooti fileTweak Hash and CRC tools as shown in Fig. 6 we retrieve the hash value of a pcap file called *Evidence One.*

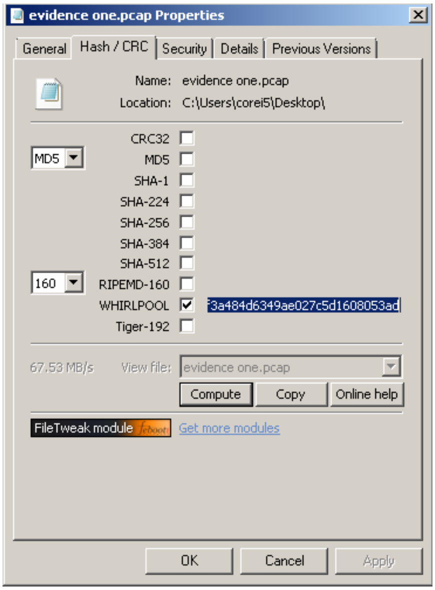
The hash value of *Evidence One's* file [original]:

```
e9a5670a95e2ed252118024acee4d7aa1972cf504970bad82cb3669c7
9fb657dcfe576606dedeace36dc16d76b8e49c145c6f3a484d6349ae0
27c5d1608053ad
```

After that, we do some modifications inside the pcap file using Hex Editor Neo and mess up contents of *Evidence One's* file. We write 0xdf at 0x0, 0xf3 at 0x1, 0xff at 0x3 with intention to ruin the information the file contained. Then we calculate again the hash value of *Evidence One* and retrieve a new hash value which is different from the earliest one.

The hash value of *Evidence One's* file [after tampering]:

```
3b63c0032fd6102dd9e0e70cefd6c7ac1e6e9f0ec1b54dbfdb6227029
3cded7024d5953e607e292d77686d241189c33826cbfbca3e49af0740
4f5920a2d12df6
```



Name: evidence one.pcap
Location: C:\Users\corei5\Desktop\

MD5 ▾
160 ▾

CRC32 □
MD5 □
SHA-1 □
SHA-224 □
SHA-256 □
SHA-384 □
SHA-512 □
RIPEMD-160 □
WHIRLPOOL ☑ f3a484d6349ae027c5d1608053ad
Tiger-192 □

67.53 MB/s View file: evidence one.pcap

Fig. 6. Retrieving hash value using Febooti fileTweek Hash and CRC

This simple experiment is to show the important of bringing reliable evidence in Courts. As we can see, how easily the changes in content can affect the hash value and would make the file inadmissible in Courts as they will question the evidence reliability. However, in this secure model, while comparing the Collection database and Presentation Database, we can conclude if the evidence was altered or not and present strong evidence to Courts.

4 Conclusions

Network Forensics is emerging as a new discipline to extract information and retrieve potential evidence from the network traffic so that, the criminals can be apprehended for the crimes they have conducted. However, as new technologies emerge, the work of forensics investigators especially the law enforcement to catch offenders in cyberspace is becoming much more complicated. Thus each evidence found is very important and needs to be secured in a legal approve ways because we don't want the criminals to escape, just because, the evidence is not reliable in courts.

In this paper, we have already presented the secure storage mechanism to secure the evidences that were collected from network traffic and display the importance of using hashing mechanism to prove the integrity of the evidences. However, we do realize that, the model is not sufficient yet to answer Jansen, W.'s [42] questions regarding the chain-of-custody and we see this as an important factor in our secure model. Who collected it? How and where? Who took possession of it? How was it stored and protected in storage? Who took it out of storage and why? [42]. Answering these questions in our future work, will append another mechanism to the model which covers access control. We would also like to enhance the capability of this model to interact with several investigators from different division assuming they were investigating the same case and using the same evidence. At the end, we hope to build working tools that can be used and implemented in real life investigations.

Acknowledgments. We would like to thank the members of Security Research Group (SRG), USM for their helpful discussions and suggestions. This work was supported by Research University grant No.1001/PKOMP/817048, School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia.

References

1. Garfinkel, S.L.: Digital forensics research: The next 10 years. Digital Investigation 73, S64–S73 (2010)
2. Hartley, W.M.: Current and Future Threats to Digital Forensics. ISSA Journal, 12–14 (2007)
3. Nance, K., Hay, B., Bishop, M.: Digital Forensics: Defining a Research Agenda. In: Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1–6 (2009)
4. Hu, L., Tang, K., Shi, G., Zhao, K.: DDCFS: A Distributed Dynamic Computer Forensic System Based on Network. In: Second International Conference on Intelligent Computation Technology and Automation, pp. 53–56 (2009)
5. Blaze, M.: Key Management in an Encrypting File System. In: Proceedings of the Summer USENIX Conference, pp. 27–35 (1994)
6. Almulhem, A.: Network forensics: Notions and challenges. In: IEEE International Symposium on Signal Processing and Information Technology, pp. 463–466 (2009)
7. Oppliger, R., Rytz, R.: Digital Evidence: Dream and reality. IEEE Security & Privacy Magazine, 44–48 (2003)
8. Selamat, S.R., Yusof, R., Sahib, S.: Mapping Process of Digital Forensic Investigation Framework. Journal of Computer Science 8, 163–169 (2009)

9. Yan, Z., Ying, L.: Research on the Key Technology of Secure Computer Forensics. In: Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 649–652 (2010)
10. Grobler, T., Louwrens, C.P., Von Solms, S.H.: A Multi-component View of Digital Forensics. In: International Conference on Availability, Reliability and Security, pp. 647–652 (2010)
11. Ho, V., Dehghantanha, A., Shanmugam, K.: A Guideline to Enforce Data Protection and Privacy Digital Laws in Malaysia. In: Second International Conference on Computer Research and Development, pp. 3–6 (2010)
12. Kent, K., Chevalier, S., Grance, T., Dang, H.: Guide to Integrating Forensic Techniques into Incident Response. NIST Special Publication 800-86. Computer Security (2006)
13. Kaushik, A.K., Pilli, E.S., Joshi, R.C.: Network forensic system for port scanning attack. In: IEEE 2nd International Advance Computing Conference, pp. 310–315 (2010)
14. Hosmer, C.: Proving the Integrity of Digital Evidence with Time. International Journal of Digital Evidence 1, 1–7 (2002)
15. Shanmugasundaram, K., Memon, N., Savant, A., Bronnimann, H.: ForNet: A Distributed Forensics Network. In: Proceedings of the Second International Workshop on Mathematical Methods, Models and Architectures for Computer Network Security, pp. 1–6 (2003)
16. Shmueli, E., Vaisenberg, R., Elovici, Y., Glezer, C.: Database Encryption - An Overview of Contemporary Challenges and Design Considerations. ACM SIGMOD Record 38, 29–34 (2009)
17. Nikkel, B.J.: Generalizing sources of live network evidence. Digital Investigation 2, 193–200 (2005)
18. Davis, M., Manes, G., Shenoi, S.: A Network-Based Architecture for Storing Digital Evidence. In: Pollitt, M., Shenoi, S. (eds.) Advances in Digital Forensics IFIP International Federation for Information Processing, vol. 194, pp. 33–42. Springer, Heidelberg (2005)
19. Beebe, N., Clark, J.: Dealing with Terabyte Data Sets in Digital Investigations. In: Pollitt, M., Shenoi, S. (eds.) Advances in Digital Forensics IFIP International Federation for Information Processing, vol, vol. 194, pp. 3–16. Springer, Boston (2005)
20. Nikkel, B.J.: Improving evidence acquisition from live network sources. Digital Investigation 3, 89–96 (2006)
21. Mahalingam, P., Jayaprakash, N., Karthikeyan, S.: Enhanced Data Security Framework for Storage Area Networks. In: Second International Conference on Environmental and Computer Science, pp. 105–110 (2009)
22. Riedel, E., Kallahalla, M., Swaminathan, R.: A framework for evaluating storage system security. In: Proceedings of the 1st Conference on File and Storage Technologies, pp. 1–16 (2002)
23. Arona, A., Bruschi, D., Rosti, E.: Adding Availability to Log Services of Untrusted Machines. In: Computer Security Applications Conference, ACSAC 1999, vol. 15, pp. 199–206 (1999)
24. Silberschatz, A., Korth, H., Sudarshan, S.: Database System Concepts. McGraw-Hill Higher Education, New York (2011)
25. Sommer, P.: The challenges of large computer evidence cases. Digital Investigation 1(1), 16–17 (2004)
26. Ren, W., Jin, H.: Modeling the Network Forensics Behaviors. In: Workshop of the 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks, pp. 1–8 (2005)

27. Kozushko, H. Digital Evidence, Graduate Seminar (2003), http://infohost.nmt.edu/~sfs/-Students/HarleyKozushko/Presentations/DigitalEvidence.pdf
28. Casey, E.: Error, Uncertainty, and Loss in Digital Evidence. International Journal of Digital Evidence 1 (2002)
29. Accorsi, R.: Log Data as Digital Evidence: What Secure Logging Protocols Have to Offer? In: 33rd Annual IEEE International Computer Software and Applications Conference, pp. 398–403 (2009)
30. Richter, J., Kuntze, N., Rudolph, C.: Securing Digital Evidence. In: Fifth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 119–130 (2010)
31. Sommer, P.: Intrusion detection systems as evidence. Computer Networks 31, 2477–2487 (1999)
32. Danielsson, J., Morch, K.H.T., Roe, P.: A system for collection and analysis of forensic evidence, GEM/05/02. Project No: 802022.Norwegian Computing Center/ Applied Research and Development (2003)
33. Scalet, S.D.: How to Keep a Digital Chain of Custody, http://www.csoonline.com/article/-220718/how-to-keep-a-digital-chain-of-custody
34. Roussev, V.: Hashing and Data Fingerprinting in Digital Forensics. IEEE Security & Privacy Magazine, 49–55 (2009)
35. Steve Friedl's Unixwix.net Tech Tips – An Illustrated Guide to Cryptographic Hases, http://www.unixwiz.net/techtips/iguide-crypto-hashes.html
36. Casey, E.: Network traffic as a source of evidence: tool strengths, weaknesses, and future needs. Digital Investigation, 28–43 (2004)
37. Cosic, J., Baca, M.: Do we have full control over integrity in Digital Evidence Life Cycle? In: Proceedings of the ITI 2010 32nd Int. Conf. on Information Technology Interfaces, pp. 429–434 (2010)
38. Kadhem, H., Amagasa, T., Kitagawa, H.: Encryption over semi-trusted database. In: Chen, L., Liu, C., Liu, Q., Deng, K. (eds.) DASFAA 2009. LNCS, vol. 5667, pp. 358–362. Springer, Heidelberg (2009)
39. Kim, H., Lee, S., Lim, J.: Digitalevidence Integrated Management System. In: Second Australian Computer, Network & Information Forensics Conference, pp. 31–39 (2004)
40. Barreto, P.S.L.M., Rijmen, V.: The Whirpool Hashing Function. In: Proceedings of First open NESSIE Workshop, pp. 1–20 (2000)
41. Une, M., Kanda, M.: Year 2010 Issues on Cryptographic Algorithm. IMES Discussion Paper Series 2006-E-8 (2006)
42. Jansen, W., Scarfone, K.: Guidelines on Cell Phone and PDA Security, NIST Special Publication 800-124. Computer Security (2008)

Attack Intention Analysis Model for Network Forensics

M. Rasmi and Aman Jantan

School of Computer Sciences, University Sciences Malaysia
Penang, 11800, Malaysia
mr77mr@hotmail.com, aman@cs.usm.my

Abstract. In network forensics, attack intentions analyses play a major role to help and accelerate decision–making for apprehending the real perpetrator. In fact, attack intention analysis is a prediction factor to help investigators to conclude a case with high accuracy. However, current techniques in attack intention analysis only focus on recognizing an alert correlation for certain evidence and predicting future attacks. In reality, more prediction factors should be used by the investigators to come to a more concise decision such as attack intention, incident path …, etc. This paper will propose an attack intention analysis model, which focus on reasoning of attacks under uncertainty intention. A new model will be introduced using a combination of a mathematical Dempster–Shafer (D-S) evidence theory with a probabilistic technique through a causal network to predict an attack intention. We found that by analyzing the attacker's intention, forensic investigation agents will be able to audit and perform evidence in an efficient way. Experiments were performed on samples of probability of attack intentions to evaluate the proposed model. Arguably, attack intention analysis model may produce a clear and impact factor for investigator decision–making.

Keywords: attacks intention, network forensics investigation, D-S theory, causal network.

1 Introduction

Network forensic techniques enable investigators to trace attacks back to the attackers. The ultimate goal is to provide sufficient evidence to prosecute the perpetrator of the crime [1]. A network forensic system can be classified according to one of three characteristics. It can be a 1) Purpose system, such as General Network Forensics (GNF) and Strict Network Forensics (SNF), 2) Complement system, such as Catch-it-as-you-can and Stop-look-and-listen, or 3) Nature system, consisting of hardware combined with software or software alone [2].

Currently, network forensics has many limitations, with specific research gaps in all areas. The major challenges are in analyzing data [3], including the variety of data sources, data granularity, data integrity, data as legal evidence and privacy issues. These challenges can be placed in three broad categories: technical, legal and resource [4]. Data analysis attempts to reconstruct attack behaviors full malicious behavior in order to understand the attacker's intention. Otherwise, the classification and clustering of network events that may arise would be more difficult [2].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 403–411, 2011.

Intention analysis often involves investigating and collecting evidence to reduce the forensic challenges [5]. However, investigation is a complex process and costly when all stakeholders and security tools are combined to enhance the investigation phase [3, 6]. Some of the more skillful attackers try to hide evidence, thus circumventing attack prevention tools. Successful attacks can exploit a series of vulnerabilities [7], as evidenced by the rapid growth of the Common Vulnerabilities and Exposures (CVE) dictionary which includes a variety of publicly known information security vulnerabilities and exposures. Most important attack information can be deduced from the CVE database, and this may provide useful evidence in the future.

Although the investigation phase complex, and it is difficult to define attacker and apprehend the perpetrator [3], mapping activities in a network forensics framework can improve and optimize the results of investigation. But it must be kept in mind that digital evidence that is obtained from the investigation phase must be acceptable in the court of law [8]. For that purpose, all security and technical roles (user, software …) in any organization should collaborate to catch the attacker and obtain its restitution. Casey [6] shows how a case study can present a broad picture of the crime that can help law enforcement agencies apprehend the perpetrator.

Attack intention is successful when a potential attack can be predicted and the goal of an attacker can be understood [7]. It is difficult for a human expert to foresee methods of intrusion [5]. An attacker will proceed to reach his goal through sequence of logical steps, using tools to hide and camouflage his patterns. Changing attack patterns is a major challenge for network forensics. The variety of network environments with a big number of attack methods makes pattern recognition more difficult [5]. False positive and false negative as an examples are the main problems in IDS, especially in misuse-based and anomaly-based detection [5]. To conclude, the limitations of security sensors and network monitoring tools make attack observation inaccurate and incomprehensive [9].

Causal networks are defined in the form of a causal poly tree (between any two nodes there exist no more than two paths) observable quantities, and unknown parameters or hypotheses [9, 23]. A causal network is a Bayesian network with an explicit requirement that the relationships be causal. Thereby, it can predict the impact of external interventions from data obtained prior to intervention in order to perform diagnosis of attack intention. The Dempster–Shafer (D-S) theory is a mathematical theory of evidence that generalizes from Bayesian probability theory [10, 22]. D-S theory helps to identify epistemic probabilities or degrees of belief. It takes the rule for combined evidence from different sources represented by a belief function, which takes into account all the available evidence. Causal networks are often used for this research where D-S theory would be more relevant [11].

This paper, presents a set of processes, shown in Fig. 1 that apply D-S theory based casual network to identify an uncertain attack intention. The proposed model will be described in section 3, where we define the equations and hypotheses depending on the collected evidence. Otherwise, the uncertain intention will be predicted from different sources of attack evidence. In the next section, we will explain the views and important ideas regarding research and different approaches. However, to determine a relationship with a proposal and explain the importance of attack analysis that needed in order to predict the intention of the attacker in network forensics. Finally, experiments and the discussion are undertaken using samples of probability of attack intentions to evaluate this model.

2 Related Works

Security analysis aims at recognizing attack plans. Attack plan analysis aims to reconstruct the scenario to find the attack intention based on a graph algorithm, with methods for intrusive intention recognition based previous studies [7]. Attack recognition is received increased interest and as main research area in artificial intelligence. Progress continues in the network security domain, and intention analysis has become an important research topic [7, 9].

Several related technologies can be connected with network forensics [2-5]. Intrusion Detection Systems (IDSs) is consists of a sensor and alarm to collect useful information for analyzing future processes. However, reliability and lack of information are its main limitations. Honeypots could be used in network forensics for studying and capturing an attacker's tools, in addition to determining the intention of the attack. The main problem with these techniques is how easily they can be attacked and compromised. Therefore, innovative methods and techniques are needed for data analysis to produce useful information, to help investigators in decision making [3].

Main intrusive intentions of an intruder, such as DoS on the web server of a host, gaining root privilege of a host, and compromising the database of a host, could be observed their behavior through 1) observing time, 2) launch host, 3) target host, and 4) rules such as intruder preconditions, network preconditions, intruder effects and network effects [12].

Proposed taxonomies to determine attack intentions depend on the goals, which use D-S evidence theory as mentioned elsewhere [13]. This technique follows the stages of attack and determines the target. The taxonomy places attacks into two categories; firstly, consequence-based attacks such as increased access, disclosure of information and denial of services, and secondly, target-based such as computer, network or files. In this research, a comprehensive view for attack intention is emerging. The studies combine each intruder state like capability, skills and tools, with system states like interest and opportunity. As a result, they determine the intention list, attack likelihood and threat estimate. D-S evidence theory is also used to investigate new traffic incident pattern-recognition approaches [14]. They used D-S theory to combine Multiple multi-class Probability Support Vector Machines (MPSVM) to the data set from different sources, to improve the accuracy and robustness of fault diagnosis.

Probabilistic approaches [9] correlate and analyze attacks, and DARPA's Grand Challenge Problem (GCP) was used to evaluate this approach. In addition, they are using this data set for identifying strategies, correlating isolated scenarios and predicting future attacks. This approach uses a probabilistic-based reasoning methods and statistical analysis for attack step correlation, and focuses on minimizing the damages in the system through developing an algorithm to correlate attack scenario in the low level correlation analysis. The main aspect of this approach is an attack tree presented from a predefined library of attacks.

Practically, an attack scenario should have a hierarchical architecture [9]. However, if an attack tree is used to define attack libraries, then attack trees can be converted into the causal network by assigning a probability of evidence and the likelihood of attack intentions. Attack tree analysis predicts a set of attack libraries represented by the graph in a similar was to attack graphs. Usually it is done manually, and is very time consuming. Otherwise, alternative techniques use the checking-model as automatic graphs, in order to construct the attack. Using model checking is an effective means to perform e- Transactions and build customer trust and confidence [15]. Strengths of this modeling type provide an efficient evaluation of protocols and are more robust than other techniques such as simulation or theorem, but has a limitations in scalability [9].

3 Attack Intention Process Model

Prediction of attack intentions depends on the nature of attack, which had been detected with its evidence. Detecting attacks depends on many security sensors and detection system products (either, commercial or non-commercial security products, such as IDS or sniffer). Knowing that, a specific attack that occurs depends on the accuracy ratio for these products. We believe the current proposed process model in this research that the attack was defined and detected with an acceptable degree of accuracy.

Fig.1. shows set of processes to select attack intention. The first process determines certain attack and ensures that attack type has been detected as well as predefined. Attack features should be constructed to find the correlation with collected evidence. This process is predefined before starting analysis of attack intention. Preparation, detection, collection and examination of evidence processes in general network forensics should be ready and predefined. For example, using the attack library, as mentioned in [9], with taking in the account that there is not a complete library for all possible attack strategy in network security, in order, to support clear and meaningful information for process 1, 2 and 3 in our proposed research model.

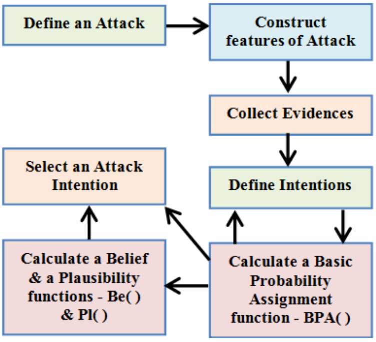

Define an Attack
Construct features of Attack
Collect Evidences
Define Intentions
Select an Attack Intention
Calculate a Basic Probability Assignment function - BPA( )
Calculate a Belief & a Plausibility functions - Be( ) & Pl( )

Fig. 1. Process Model Based on D-S evidence for Attack Intention

Preliminary hypotheses, prior probability of parent node's (attack) and collecting of evidence are major factors to apply next process. A hypothesis depends on several elements, such as accuracy of attack detection, evidence collection and the initial potential of intention. Attack intention, in many cases, is related to attacking goals and the volume of damages that proceed from the attack. Using probabilistic laws in the causal network and according to the evidence analysis, probability for each of evidence will be computed. Depending on these results, probability of each intention with given conditional evidence will be computed too.

Basic Probability Assignment (BPA) that is calculated for a set of mutual intentions to present the strength of some evidence (not attack evidence) for this intention. BPA function also known as an exact belief in the proposition represented by intention. From BPA, the belief and plausibility probability can be defined. Finally, all believes values will be compared in order to select the highest one that will be an attack intention.

This model represents a solution in causal network using D-S evidence theory, in order to find a potential of attack intention. To implement a model, assume that the IDS, firewalls or any detection system has been used, detected an attack a_1 with a proper correctly positive value between [0.50, 0.99], this value means that attack a_1 was detected with an adequate value between [0.50, 0.99].

Accuracy of detecting an attack for a_1 type represented from the same positive value, and these are the first hypotheses in this model. Evidence related to this attack was collected and reserved from previous phases of network forensics, using suitable frameworks and tools. Believe and prediction of uncertain attack intention depends on available information, initial proper values and hypotheses. Attack a1 has conditional dependent evidence $\{ev_1, ev_2, ev_3 \ldots ev_n\}$.

Assume that, there is a set of uncertain attack intentions $\{i_1, i_2, i_3 \ldots, i_n\}$, each one is connected and dependent on at least one evidence. Each intention has a very low and unimportant prior probability between [0.001, 006] this value doesn't effect for prediction of the real intention. Even though, it's legally known that a suspect is always innocent until proven guilty. Prior probability for attack intention represented this general base in the judicial life system, which means that the attack intention is close to undefined value in the initialization step. For example, in (0.85) accuracy, if we suppose that an attack a_1 is detected from IDS and there is a set of evidences $\{ev_1, ev_2, ev_3, ev_4, ev_5, ev_6\}$. The intention of a1 that depends on the previous evidences has a set of intentions $\{i_1, i_2, i_3\}$. Depending on the previous information, we can assume that:

The prior probability of i_1=i_2=i_3=0.005, then we can claim that the a_1 intention was most unlikely to be i_1, i_2 or i_3. In other words, the a_1 intention is not i_1 (the complement of $P(i_1) = P(i_2) = P(i_3) = 1-0.005=0.995$). The proposed model supposes that each intention related and connected to attack evidence. For example, if we have the following set of evidences $\{ev_1, ev_3, ev_5\}$ then the attack intention will be i_1. This stage depends on the evidence analysis and intention prediction. We can say now if we found $\{ev_1, ev_3, ev_5\}$ for a_1, then the probability that there is an ev_1 where given an i_1 is 0.85 (after using network forensics frameworks and tools).

Table 1 shows the sample of prior probability for this model to find a probability of attack a_1 intention i_1. The purpose of the Attack Intention Analysis model is to answer the following question: "what is the probability of actual i_1, when a_1 is detected with a set of evidences $\{ev_1, ev_3, ev_5\}$?"

Table 1. The prior's probability of attack a_1 intention i_1

Probability	Value	Description
$P(a_1)$	0.85	Attack a_1 was detected. (Depends on accuracy ratio of detection system)
$P(\neg a_1)$	0.15	Doubts ratio, that a_1 doesn't detected
$P(i_1)$	0.005	The prior probability of i_1 as actual intention for attack a_1
$P(\neg i_1)$	0.995	The probability that i_1 not intention for attack a_1
$P(ev_1 \mid i_1)$	0.85	The probability that is there an ev1 where given an i_1. (Depends on evidence collection, initialize information , set of hypotheses and after using network forensics frameworks and tools)
$P(\neg ev_1 \mid i_1)$	0.15	The probability that ev_1 if a_1 intention not related to i_1.

To solve this problem, firstly, the probability of ev_1 should be calculated regards of its available information, and it's equal to the probability of ev_1 when i_1 occurs. It adds to the probability of ev_1 when the intention not related to i_1. As in the following equation:

$$P(ev_1) = P(ev_1 \mid i_1) * P(i_1) + P(\neg ev_1 \mid i_1) * P(\neg i_1) \tag{1}$$

Repeat Eq.(1) for $P(ev_3)$, $P(ev_5)$, $P(ev_1,ev_3)$, $P(ev_1,ev_5)$, $P(ev_3,ev_5)$ and $P(ev_1,ev_3,ev_5)$. After that we can compute the probability of i_1 for each given calculated probability of evidence, as mentioned above, using the following equation, for ev1 probability:

$$P(i_1 \mid ev_1) = \frac{P(ev_1 \mid i_1) * P(i_1)}{P(ev_1)} \tag{2}$$

Then, Eq.(2) is used to compute the probabilities $P(i_1|ev_3)$, $P(i_1|ev_5)$, $P(i_1|\, ev_1,ev_3)$, $P(i_1|ev_1,ev_5)$, $P(i_1|ev_3,ev_5)$ and $P(i_1|ev_1,ev_3,ev_5)$. In the result, we can say that the probability of an intention i_1 for attack a_1 with the evidences $\{ev_1, ev_3, ev_5\}$ is the output of the following equation:

$$P(i_1 \mid ev_1, ev_3, ev_5) = \sum_{ev=1,3,5} \frac{P(ev \mid i_1) * P(i_1)}{P(ev)} \tag{3}$$

D-S theory starts by assuming a frame of discernment, which is a set of evidence EV_s related to the uncertain intention i_1 for attack a_1. EV_s would be the set consisting of all possible evidences for i_1 as described above. Elements of 2^{EV_s}, i.e. subset of EV_s, are the class of general propositions concerning the actual state of the attack a_1 evidence domain, including empty set Ø.

$$\begin{aligned} &EV_s = \{ev_1, ev_3, ev_5\} \\ &\therefore 2^{EV_s} = \{\phi, \{ev_1\}, \{ev_3\}, \{ev_5\}, \{ev_1, ev_3\} \\ &\{ev_1, ev_5\}\{ev_3, ev_5\}\{ev_1, ev_3, ev_5\}\} \end{aligned} \tag{4}$$

The proposed model assumes that $BPA(ev) = P(i_1|ev)$ for each ev ϵ 2^{EV_s}. That $P(i_1|ev)$ acts as an actual i_1 probability when a_1 is detected with a set of evidences

$\{ev_1, ev_3, ev_5\}$. The function BPA:$2^{EVs} \rightarrow [0, 1]$ is called a belief or support function Be( ), if it satisfies Be(Ø) = 0 and Be(EV_s) = 1. In other words, it is the amount of justified support to EV_s. In probabilistic formalisms, it is generally seen as the lower probability function of D-S theory. The Be(EV_s) for a set of EV_s is defined as the sum of all the BBA_s of subsets of the interests, as in the following equation:

$$\mathrm{Be}\left(\mathrm{EV_q}\right)= \sum_{\mathrm{EV_r}|\mathrm{EV_r}\subseteq\mathrm{EV_q}} \mathrm{BPA(EV_r)}\forall \mathrm{EV_r} \subseteq \mathrm{EV_s} \tag{5}$$

The plausibility Pl( ) is the amount of potential support to EV_s. It means support evidence, which is not strictly given to EV_s. In probabilistic formalisms, it is generally seen as the upper probability function of D-S theory. It is the sum of all the BPAs of the sets EV_r intersects the set of interest EV_q, as in the following equation:

$$Pl\left(EV_q\right)= \sum_{EV_r|EV_r\cap EV_q\neq\phi} BPA(EV_r)\forall EV_r \subseteq EV_s \tag{6}$$

The explanations of above equations are summarized in Table 2. (BPA(EV_q), Be(EV_q) and Pl(EV_q)), which are related to the above example of attack a_1 intention i_1, are computed, all of the general hypotheses and propositions for evidence and intentions in the domain of attack $a_1$should be declared. After determining the probability of i_1 for a_1, we repeat the same process for other a_1 intentions set $\{i_2, i_3\}$.

Table 2. BPA, Belief and plausibility functions for intention i_1

Hypotheses/Propositions	BPA	Belief	Plausibility
Null, no evidences for i_1	0	0	0
Probability of actual i_1, when a_1 detected with evidence ev_1. $P(i_1\|ev_1)$	0.03	0.03	0.82
Probability of actual i_1, when a_1 detected with evidence ev_3. $P(i_1\|ev_3)$	0.05	0.05	0.86
Probability of actual i_1, when a_1 detected with evidence ev_5. $P(i_1\|ev_5)$	0.04	0.04	0.84
Probability of actual i_1, when a_1 detected with evidence ev_1,ev_3. $P(i_1\|ev_1,ev_3)$	0.08	0.16	0.96
Probability of actual i_1, when a_1 detected with evidence ev_1,ev_5. $P(i_1\|ev_1,ev_5)$	0.07	0.14	0.95
Probability of actual i_1, when a_1 detected with evidence ev_3,ev_5. $P(i_1\|ev_3,ev_5)$	0.09	0.18	0.97
Either or any evidences (probability of actual i_1, when a_1 detected with evidence ev_1,ev_3,ev_5. $P(i_1\|ev_1,ev_3,ev_5)$	0.64	1	1

4 Experimental Results and Analysis

Without evidence, there is no way to determine attack attention or provide a solution to it. This is prove, as shown in Table 2, by setting the first hypothesis or null hypothesis to true and zero. Based on the other hypothesis for intention 1, we found that Pl( ) ≥ Be( ) and Be( ) ≥ BPA( ) for each hypothesis. This means that any potential support for evidence is always greater than or equal to justified support.

If we collect a single piece of evidence for each intention, the probability of belief is an actual attack intention that will not be accurate, in contrast to combining it with another type of evidence, for example, the belief probability of actual i_1, when the attack detected with evidence ev_1 and ev_5 (=0.14) is greater than the belief probability of actual i_1, when the attack detected with evidence ev_5 (=0.04). An attack intention that has more evidence has an increased probability of being an actual attack. In the result, the accuracy of this intention will be increased.

The accuracy of prediction for any intention related to the amount of evidence collections and the strength related with the intention. Number of evidence with their detection accuracy effect on the BPA values that means if we assume that the intention related with five evidence, and we detect only four, in particular, accuracy. The accuracy will be lower than if we detected all the five evidence.

5 Conclusions and Future Work

This research, propose a new model to predict attack intention with more accuracy using D-S evidence theory combined with causal networks. The proposed model will facilitate the choices for framing a decision making, in order to obtain clear information and achieve acceleration of the investigation phase. Network forensic investigators can use this model to enclose all the potential forestation of attack intention, and choose the best one to take actions.

Attack analysis is a critical and challenging task in the network forensics management. Furthermore, intention recognition and analysis are an important research area in artificial intelligence and security fields. Obviously, to gain high probability for attack intention, we need deep analysis and more efforts to prepare evidence using suitable network security tools.

Acknowledgments

This work was supported by RU Grant No.1001/PKOMP/817048, School of Computer Science, University Sciences Malaysia, Penang, Malaysia.

References

1. Yasinsac, A., Honeytraps, M.Y.: A network forensic tool. In: Proceedings of the Sixth Multi-Conference on Systemics, Florida, USA (2002)
2. Pilli, E.S., Joshi, R.C., Niyogi, R.: Network forensic frameworks: Survey and research challenges. Digital Investigation (2010) (in Press, Corrected Proof)
3. Almulhem, A.: Network forensics: Notions and challenges. In: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT (2009)
4. Rogers, M.K., Seigfried, K.: The future of computer forensics: a needs analysis survey. Computers & Security 23(1), 12–16 (2004)
5. Huang, M.-Y., Jasper, R.J., Wicks, T.M.: A large scale distributed intrusion detection framework based on attack strategy analysis. Computer Networks 31(23-24), 2465–2475 (1999)
6. Casey, E.: Case study: Network intrusion investigation - lessons in forensic preparation. Digital Investigation 2(4), 254–260 (2005)

7. Peng, W., Yao, S., Chen, J.: Recognizing Intrusive Intention and Assessing Threat Based on Attack Path Analysis. In: International Conference on Multimedia Information Networking and Security, MINES (2009)
8. Siti Rahayu Selamat, R.Y., Sahib, S.: Mapping Process of Digital Forensic Investigation Framework. IJCSNS International Journal of Computer Science and Network Security 8(10), 163–169 (2008)
9. Qin, X., Lee, W.: Attack plan recognition and prediction using causal networks. In: 20th Annual Computer Security Applications Conference (2004)
10. Shafer, G.: A Mathematical Theory of Evidence, 1st edn. Princeton University Press, Princeton (1976)
11. Burrus, N., Lesage, D.: Theory of Evidence (DRAFT), Technical Report. Le Kremlin-Bicêtre CEDEX, France (2003)
12. Wang, Z., Peng, W.: An Intrusive Intention Recognition Model Based on Network Security States Graph. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCom 2009 (2009)
13. Peng, W., et al.: Recognizing Intrusive Intention Based on Dynamic Bayesian Networks. In: International Symposium on Information Engineering and Electronic Commerce, IEEC 2009(2009)
14. Zeng, D., Xu, J., Xu, G.: Data Fusion for Traffic Incident Detector Using D-S Evidence Theory with Probabilistic SVMs. Journal of Computers 3(10), 36–43 (2008)
15. Bonnie Brinton, A., et al.: The application of model checking for securing e-commerce transactions. Commun. ACM 49(6), 97–101 (2006)
16. Popescu, D.E., Lonea, M., Zmaranda, D., Vancea, C., Tiurbe, C.: Some Aspects about Vagueness & Imprecision in Computer Network Fault-Tree Analysis. International Journal of Computers Communications & Control 5(4), 558–566 (2010); ISSN 1841-9836
17. Reiz, B., Csató, L.: Bayesian Network Classifier for Medical Data Analysis. International Journal of Computers Communications & Control 4(1), 65–72 (2009); ISSN 1841-9836

Formal Verification for Interaction Protocol in Agent-Based E-Learning System Using Model Checking Toolkit - MCMAS

Norizal Abd Latif, Mohd Fadzil Hassan, and Mohd Hilmi Hasan

Department of Computer and Information Sciences,
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
`eja.norizal@gmail.com,`
`{mfadzil_hassan,mhilmi_hasan}@petronas.com.my`

Abstract. This paper presents a formal verification done for interaction protocol in agent-based e-learning system using a model checking toolkit – MCMAS (Model Checking Multi-Agent System). The goal of this interaction protocol is to automate the document downloading and notification in e-learning on behalf of the students. The specification of the interaction protocol for each agents are translated into Interpreted Systems Programming Language (ISPL) file as the programming language for MCMAS and a combination of temporal logic operators like Computation Tree Logic (CTL) and Linear Temporal Logic (LTL) are used to verify the formula for each tested reachable property. The purpose of executing this formal verification is to convince that this interaction protocol is sound and reachable for each state. Overall, this paper describes the idea of agent-based e-learning system, MCMAS toolkit - used as a model checker, CTL and LTL – logics in verification used, ISPL – programming language used under MCMAS platform and the results derived from the running verification.

Keywords: Agent-based, MCMAS, Interaction Protocol, ISPL, Logics, Model Checking, Verification, Reachability.

1 Introduction

Relying on e-learning system has become a trend to the current education field specifically in institutions of higher learning (IHLs) to support their learning system [1]. This is coherent with the rapid advancement in current technology where many multimedia components like live lecture webcast from other regions, video streaming, interactive lecture slides, online quizzes and tests are embedded in this e-learning system. Taking this kind of technology advantages, e-learning system is now able to integrate with another new emerging component of technology which is agent-based technology that is widely used mainly in intelligent tutoring system (ITS) [2, 3] .

Conventionally, e-learning is used as a repository for storing teaching materials like lecture notes and slides which is only capable on one-way communication where the e-learning system behaves as a receiving party by not responding to the stored documents actively to the main user of this e-learning system; the student. This is

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 412–426, 2011.

based on what had happened when the lecturers upload the lecture notes, for instance, into the e-learning. The e-learning system will post new notification to the students profile informing the newly lecture notes has been uploaded by their lecturers. This is a common way of notification in any e-learning system.

But, by adopting the agent-based technology, e-learning system now has become more intelligent by responding to whatever happened in its system to any communicating parties, the students and also the lecturers. For example, if there is newly uploaded lecture notes, the agents resides in e-learning system, will download the lecture notes on behalf of the students to avoid them from missing any new updates from their lecturers either the cancellation class announcement, updated lecture notes or any online quizzes conducted by their lecturers.

In accordance to this scenario, a study on agent interaction protocol (AIP) is done and specifications of each agent are specified. Upon the completion of this specification, a formal verification for each protocols are verified using a model checking toolkit named as Model Checking Multi-Agent System (MCMAS) where the purpose of having this verification is to check that each states within this protocols are reachable. In short, a reachability property is checked using the temporal logic of CTL and LTL.

The objective of this paper is to describe a formal verification done using model checking toolkit – MCMAS. The fundamental of formal verification is also described together with short details on temporal logics used for the verifications. The derived results are also described in this paper.

The Introduction part describing the overall idea of this paper is mentioned briefly in Section 1. Section 2 on the other hand, describes the background study of this research mentioning why this study is done. In Section 3, the fundamental components of formal verification is described in detail starting from the definition of formal verification, model checking approach, the process of model checking, temporal logics (Kripke structure, Computation Tree Logic - CTL, Linear Temporal Logic - LTL) and a brief explanation on MCMAS toolkit. Section 5, shows the results derived from MCMAS toolkit and Section 6 concludes the paper based on the results derived.

2 Background Study

Most companies and institution of higher learning (IHLs) nowadays are using e-learning system as a medium to communicate with their employees or students. However, the focus of this study is more on the higher learning institutions level where some universities rely heavily on e-learning system to support their learning system. The companionship of e-learning to the students does assist them in learning process nowadays since many multimedia components like live webcast from foreign lecturer, interactive lecture notes, online quizzes even digital library has been embedded in e-learning system which enhance the comprehension of the students to learn certain subjects.

With the rapidness of current technology advancement nowadays, an agent-based e-learning system was able to be introduced. It is called Intelligent Tutoring System (ITS) which it assists students who needs help to learn a particular subject at any time they want. This example proves that the application of software agent in e-learning system can be implemented mainly in e-learning domain.

Although there is multimedia widgets that can be embedded in e-learning system, students still facing a hassle to re-login to the e-learning system every time they want to know what is the latest information posted by their lecturers. This scenario is prone to a situation where sometimes students will forget to check their e-learning announcement regularly and they will miss the important announcement like an announcement of class cancellation or change of venue for the scheduled class. To avoid this situation happened, an idea to adopt an agent-based for e-learning system is considered as a solution for the students benefit.

By adopting the agent-based system in e-learning, all the documents uploaded by their lecturers is now will be downloaded by an agent on behalf of the students based on the specification protocol specified for each agents involved. To accomplish this idea, there are three types of agents are identified to execute this goal which is Lecturer Agent, Student Agent and E-learning agent. These three agents will communicate between each other in order to accomplish their goal which is to automatically download any documents uploaded by the Lecturer Agent into student's desktop. By doing this, all the students will not missing the important information anymore.

This is only the overall idea of how a specification of agent interaction protocol will be formulated. This is because in this research area, the interaction protocol will be studied first before any agent-based development is implemented. Hence, the focus of this study is to develop a specification for agents in e-learning to do an automatically document downloading and notification on behalf of the students. Once the specification is specified, a formal verification is done based on model checking approach using MCMAS to check the reachability property for each states.

3 Fundamental Components for Formal Verification

Before running any verification, some understanding on relevant foundations is required. Those foundations includes the definition of verification itself, a basic understanding on temporal logics which comprises of CTL and LTL, model checking technique also need to be grasped together with its model – Kripke structure. Then, a suitable toolkit to run the verification like MCMAS tool needs to be understood as well. Below is some basic description on the related components.

3.1 Formal Verification

Verification can be defined as a process of verifying a system whether it satisfies its design requirements or not [3]. Verification is also crucial to avoid unwanted behaviors. The most commonly technique for verification is testing and simulation [4]. This verification technique requires a number of test cases which might not able to discover any deadlock situations [4]. Hence, appealing alternative to simulation and testing is the approach of formal verification, which is a class of logic-based techniques.

As per described by [4], the difference between formal verification and testing or simulation is formal verification conducts an exhaustive exploration of all possible behaviors while testing and simulation explore some of possible behaviors. Since failure is unacceptable in any developed system, formal verification has become a reliable verification technique to check any bugs that may lurk in the system.

Since there are many techniques to do verifications like testing, simulation, theorem proving or deductive reasoning, model checking technique is chosen because a desired behavioral property like reachability property in agent-based e-learning system can be verified over a given system using a model through exhaustive enumeration of all states reachable by the system and also the behaviors that traverse through the states [5, 6].

3.2 Model Checking

Model checking can be defined as a technique for verifying finite state concurrent systems [4]. Model checking technique enjoys two remarkable advantages:

- *Automatic* – Verification by model checking is performed automatically and its application requires no expertise in mathematical disciplines such as theorem proving. Hence, anyone who can run simulations of any system is fully qualified and capable to do model checking for the same system [4],
- *Counterexample* – When the system fails to satisfy a desired property, verification by model checking always produces a counterexample that demonstrates a behavior which falsifies the desired property. This faulty trace provides a priceless insight to understanding the real reason for the failure which giving hints on how to fix the problem [4].

Model checking uses an exhaustive search to determine if some specification is true or not and it will always terminate with a yes/no or true/false answer [4].

3.3 Temporal Logics

Temporal logic by definition is a form of logic specifically tailored for statements and reasoning which involve the notion of order in time [6]. Temporal logic is also a formalism for describing sequences of transition between states in a system. Temporal logics have proved to be useful for specifying concurrent systems since it is able to describe the ordering of events in time without introducing time explicitly [4].

According to [4] as well, by adopting temporal logics, time is not mentioned explicitly; instead, a formula might specify that eventually some designated state is reached, or that an error state is never entered. From [4], it mentioned that properties like eventually or never are specified using special temporal operators which can be combined with Boolean connectives that provide the semantics for the operators. The meaning of a temporal logic formula is determined with respect to a labeled state-transition graph known as Kripke structure [3, 4].

Kripke Structure. Kripke structure is used to capture behavior of system. According to [4], a Kripke structure consists of a set of states, a set of transitions between states and a function that labels each state with a set of properties that are true. Paths in a Kripke structure model computations of the system. A Kripke structure representation is described as below.

Let AP be a set of atomic propositions. A Kripke structure M, over AP is a four tuple M = (S, S0, R, L) where:

- S is a finite set of states.
- S0 ⊆ S is the set of initial states.
- R ⊆ S X S is a transition relation that must be total, which is, for every state s ∈ S there is a state of s' ∈ S such that R(s, s').

Computation Tree Logic. Computation Tree Logic (CTL) is a temporal logic that having connectives which allow to refer the future [7]. CTL formulas describe properties of computation trees. The tree is formed by designating a state in Kripke structure as the initial state and then unwinding the structure into an infinite tree with the designated state at the root. This computation tree shows all of the possible executions starting from the initial state [7]. According to [4], CTL formulas are composed of:

- *Path Quantifiers* – is used to describe the branching structure in the computation tree which consist two basic quantifiers,
- *Temporal Operators* – is used to describe the properties of a path through the tree which consist five basic operators.

Table 1. Table showing 2 basic quantifiers and 5 temporal operators

Path Quantifiers	Temporal Operators
1) **A** - "for all computation paths".	1) **X** – "next time" requires that a property holds in the second state of the path.
2) **E** – "for some computation path".	2) **F** – "eventually" or "in the future" is used to assert that a property will hold at some state on the path.

There are two types of formulas in CTL which is:

- *State Formulas* – which are true in a specific state,
- *Path Formulas* – which are true along a specific path.

The syntax of state and path formulas is based on this rule as per described in Table 2 below.

Table 2. Table showing rules of syntax for state formulas

State Formulas	Path Formulas
1) If $p \in AP$, then p is a state formula.	1) If f is a state formula, then f is also a path formula.
2) If f and g are state formulas, then $\neg f, f \lor g$ and $f \land g$ are state formula.	2) If f and g are path formulas, then $\neg$f, $f \lor g, f \land g$ **X**f, **F**f, **G**f, f**U** g and f **R** g are path formulas.

Based on [8], in summary CTL can be defined as by

$$\varphi ::= \mathrm{p} \mid \neg\, \varphi \mid \varphi \lor \varphi \mid EX\varphi \mid EG\ \varphi \mid E\ [\varphi\ U\ \varphi] \mid EF\varphi \mid AX\ \varphi \mid AG\ \varphi \mid A[\varphi\ U\ \varphi] \mid AF\varphi \mid$$

where in this definition,

- $p \in P$ is an atomic formulas,
- EX φ is read as "there exists a path such that at the next state φ holds",
- EG φ is read as "there exists a path such that φ holds globally along the path",
- E [φ U φ] is read as "there exists a path such that φ holds until ψ holds",
- $EF\varphi$ is read as "there exists a path such that φ holds at some future point",
- AX φ is read as "for all paths, in the next state φ holds",
- AG φ is read as "for all paths, φ holds globally",
- A[φ U φ] is read as "for all paths, φ holds until ψ holds",
- AFφ is read as "for all paths, φ holds at some point in the future".

CTL operators above are composed of pair of symbols where:

- *First Symbol* – is a quantifier over paths (E),
- *Second Symbol* – is expressing some constraints over paths.

Linear Temporal Logic (LTL). CTL is an important specification language for reactive system, but it is not the only language [6]. Alternative formalism includes Linear Temporal Logic (LTL) which in contrast to CTL, is a logic to reason about linear sequences of states that can be defined in terms of a set of atomic propositions P, as follows [4]:

$$\varphi ::= \mathrm{p} \mid \neg\, \varphi \mid \varphi \lor \varphi \mid X\varphi \mid \varphi\ U\ \varphi \mid \mathrm{G}\ \varphi$$

LTL is closely related to CTL since it shares similar expressive mechanisms, such as an *Until*, *U* connective. Unlike CTL, its formulas have meanings on individual computation paths, that is there is no explicit path quantifiers *E* and *A*. Hence, LTL appears less expressive that CTL, However, LTL allows one to nest Boolean connectives and modalities in a way not permitted by CTL in which it makes LTL appears more expressive [9].

Reachability Property. According to [6], reachability property states that some particular situation can be reached. In this research, all states are compulsory to be reached in order to deliver its goal. From the viewpoint of model checking, a justification has been made that reachability property is usually the most crucial in system correctness which happen to be the easiest one to check [6].

As mentioned by [6], when reachability properties are expressed in temporal logic, the EF combinatory appears naturally. Reachability properties is defined and written in this way $EF\ \varphi$, with φ a propositional formula free of temporal combinators where φ is commonly called as *a present tense* formula in the temporal logic setting. $EF\ \varphi$ is read as "there exists a path from the current state along which some state satisfies φ".

Since reachability properties are typically the easiest to verify, a model checking tool like MCMAS for instance is able to construct the reachability graph of a system, in which it can answer any reachability question by simply examining the generated graph even if it does not include temporal logic [6].

3.4 MCMAS – Model Checker Multi-Agent System Toolkit

According to [9], Model Checking Multi-Agent Systems (MCMAS) is a model checker for multi-agent systems which it is possible to verify time and knowledge. MCMAS is implemented in C++ which can be compiled for all major platforms. MCMAS also implements standard algorithms for CTL [10].

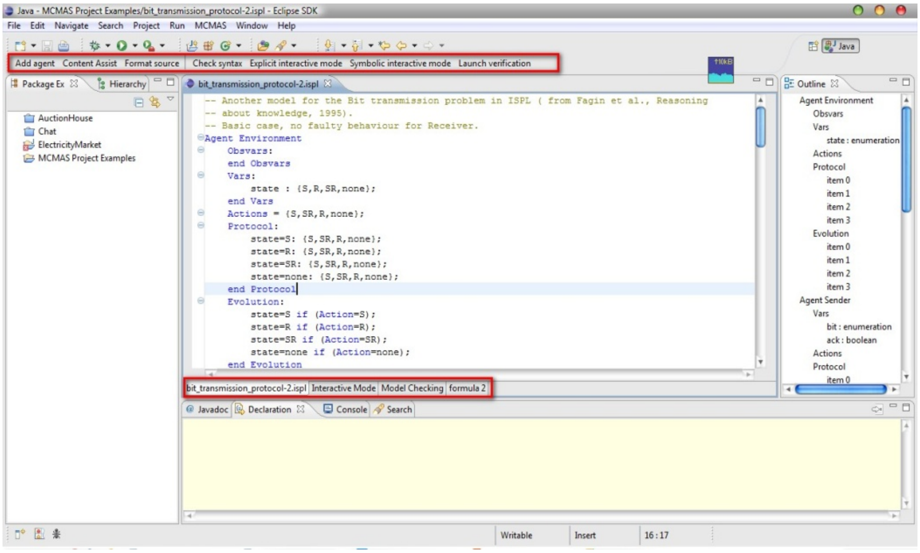
MCMAS accepts the description of a Multi-Agent System (MAS) using interpreted systems in input and a list of formulae to be verified. Since MCMAS takes ISPL descriptions as input, its ISPL file is fully describes a multi-agent system in both, the agents and the environment which is closely follows the framework of interpreted systems as described in the ISPL section [11].

MCMAS graphical user interface (GUI) is built as an Eclipse plug-in and some of functionalities are listed and illustrated in Fig. 1 below:

- *ISPL Program Editing* – guides user to create and edit ISPL program,
- *Interactive Execution Mode* – is used to explore the model.
- *Counterexample Display* – is used to launch the verification process via the GUI which calls the checker.

Interpreted Systems Programming Language - (ISPL). From [8], mentioning that interpreted systems is described in MCMAS using Interpreted Systems Programming Language (ISPL). ISPL programs are types of text files describing a MAS using a formalism of interpreted systems which includes [12]:

- A list of Agents' descriptions – is described by giving the agents possible local states, their actions, protocols and local evolution function;
- An *evaluation* function;
- A set of *initial states;*
- A list of *groups* (used for group modalities);
- A list of *formulae.*


Java - MCMAS Project Examples/bit_transmission_protocol-2.ispl - Eclipse SDK
File Edit Navigate Search Project Run MCMAS Window Help
Add agent Content Assist Format source Check syntax Explicit interactive mode Symbolic interactive mode Launch verification
Java
Package Ex
Hierarchy
AuctionHouse
Chat
ElectricityMarket
MCMAS Project Examples
bit_transmission_protocol-2.ispl
Outline
Agent Environment
Obsvars
Vars
state : enumeration
Actions
Protocol
item 0
item 1
item 2
item 3
Evolution
item 0
item 1
item 2
item 3
Agent Sender
Vars
bit : enumeration
ack : boolean
Actions
Protocol
item 0
bit_transmission_protocol-2.ispl Interactive Mode Model Checking formula 2
Javadoc Declaration Console Search
Writable Insert 16 : 17

```
-- Another model for the Bit transmission problem in ISPL ( from Fagin et al., Reasoning
-- about knowledge, 1995).
-- Basic case, no faulty behaviour for Receiver.
Agent Environment
    Obsvars:
    end Obsvars
    Vars:
        state : {S,R,SR,none};
    end Vars
    Actions = {S,SR,R,none};
    Protocol:
        state=S: {S,SR,R,none};
        state=R: {S,SR,R,none};
        state=SR: {S,SR,R,none};
        state=none: {S,SR,R,none};
    end Protocol
    Evolution:
        state=S if (Action=S);
        state=R if (Action=R);
        state=SR if (Action=SR);
        state=none if (Action=none);
    end Evolution
```

Fig. 1. MCMAS GUI Menus

4 Results and Discussion

In this section, results for each protocol is discussed and analyzed. There are four protocols in this study which comprises of upload, prioritize, broadcast and download. Below is the description of generated result for each protocol.

Upload Protocol. Upload protocol is where the user of e-learning system will use it to upload any documents into the e-learning system. This protocol is designed specifically for the Lecturer Agent since this study is all about using a software agent to work on behalf of the Lecturer itself.

Therefore, to imitate the common upload process done by human being (*in this case the Lecturer*), a similar mechanism for uploading a document is specified. The common upload process can be summarized as shown in this state diagram (Fig. 2) below.

From Fig. 2 diagram, it is clearly indicates that the common states for upload process involves a user (Lecturer) and system (E-Learning) but when involving interaction between two software agents, further states like waiting and notifying states need to be expanded as shown in Fig. 3. This is because in interaction, the communicating parties will expect to have feedback on each sending message. For example, E-Learning agent will be waiting a request from Lecturer agent to upload a document and from that initiation, the interaction will begin.

From Fig. 2 and Fig. 3, it is obvious that at least there are two main agents required to interact in order to execute this upload protocol, which is Lecturer Agent and Elearning Agent. This extended states diagram indicates the possible states that may exist during the interaction to upload a document into the e-learning system where this states is declared in the declaration section of ISPL file as shown in Fig. 5.

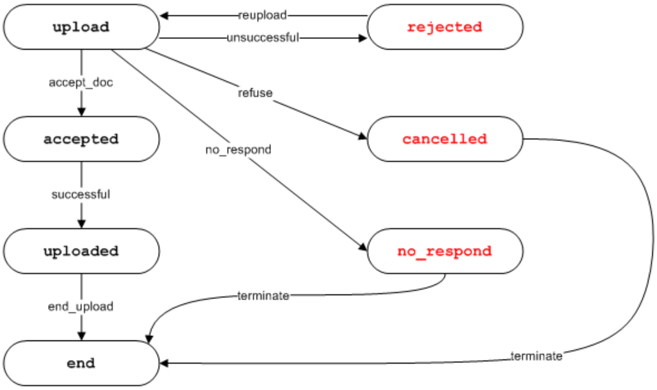


Fig. 2. State Diagram for Upload Protocol

In Fig. 5, the action that may be taken when reaching the specified states is also declared followed by the specification of the protocol in protocol section for each agents. Then, the evolution from one state to another state is also defined for both communicating agents in order to see the successive flow of upload document interaction.

An evaluation is defined in evaluation section in ISPL file where the necessary states that need to be evaluated for example to check whether the upload process is successful or not by stating which state to be reached in order to indicate the upload is a success. This evaluation statement will be referred in the formulae section where the reachability property is test using the 'EF' expression.

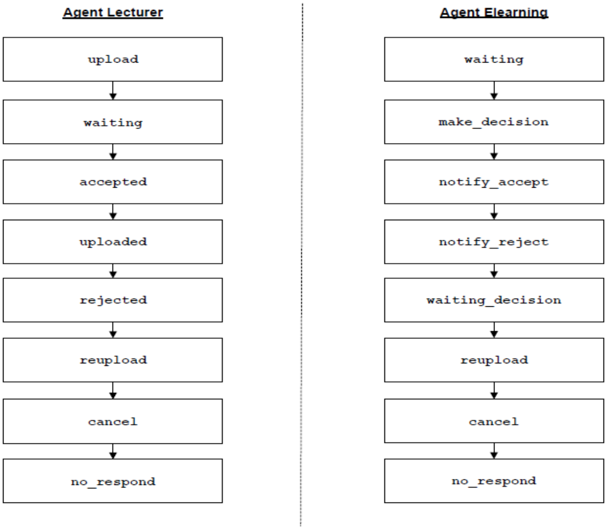
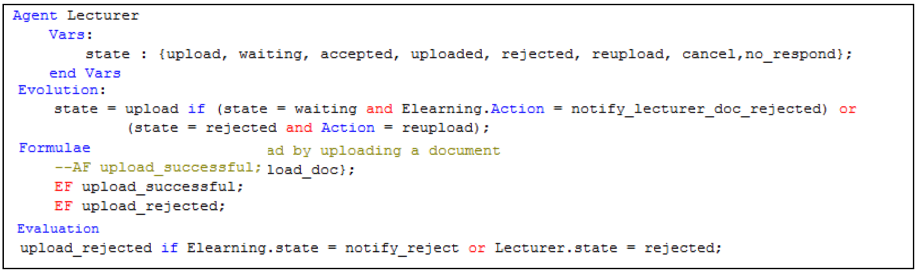


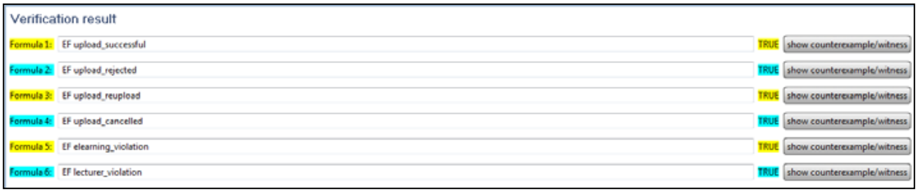
Fig. 3. Extended State Diagram for Upload Protocol

In this study, there are five reachability expressions that are evaluated using this model checking tool, MCMAS. Fig. 5 shows the verification result for upload protocol in which it shows 'True' in all expression checked indicating that all states specified in the Evaluation section earlier are reachable. The directed graph in Fig. 6 proves the states prescribed earlier are reachable.



```
Agent Lecturer
    Vars:
        state : {upload, waiting, accepted, uploaded, rejected, reupload, cancel,no_respond};
    end Vars
 Evolution:
     state = upload if (state = waiting and Elearning.Action = notify_lecturer_doc_rejected) or
             (state = rejected and Action = reupload);
 Formulae                    ad by uploading a document
     --AF upload_successful; load_doc};
     EF upload_successful;
     EF upload_rejected;
 Evaluation
 upload_rejected if Elearning.state = notify_reject or Lecturer.state = rejected;
```

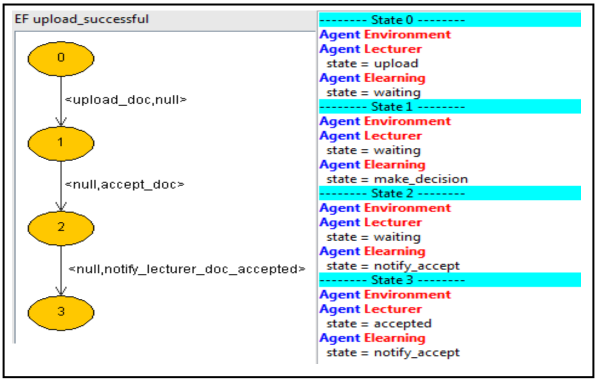
Fig. 4. Upload Protocol In ISPL File



Verification result

Formula 1:	EF upload_successful	TRUE	show counterexample/witness
Formula 2:	EF upload_rejected	TRUE	show counterexample/witness
Formula 3:	EF upload_reupload	TRUE	show counterexample/witness
Formula 4:	EF upload_cancelled	TRUE	show counterexample/witness
Formula 5:	EF elearning_violation	TRUE	show counterexample/witness
Formula 6:	EF lecturer_violation	TRUE	show counterexample/witness

Fig. 5. Verification Result For Upload Protocol



```
-------- State 0 --------
Agent Environment
Agent Lecturer
 state = upload
Agent Elearning
 state = waiting
-------- State 1 --------
Agent Environment
Agent Lecturer
 state = waiting
Agent Elearning
 state = make_decision
-------- State 2 --------
Agent Environment
Agent Lecturer
 state = waiting
Agent Elearning
 state = notify_accept
-------- State 3 --------
Agent Environment
Agent Lecturer
 state = accepted
Agent Elearning
 state = notify_accept
```

Fig. 6. All States Are Reachable In Formula 1 (*Upload Protocol*)

Prioritize Protocol. Prioritize protocol is an optional protocol since it will be activated when there is same document type is uploaded at same time by the same lecturer. Hence, a priority mechanism is required to handle which documents will be notified first to the Student agent based on the level of urgency of the document uploaded in the specific folder.

From the folder location, the priority level of the document can be identified for this study. Therefore in Fig. 7, the first thing the internal E-Learning agent will do is to get the location of the document uploaded by the Lecturer in order to start the prioritizing task. After acquiring the information for the document location, the priority assignment will be executed and the document with priority flag will be sorted according to its priority number. In this study, the highest priority (1st priority) is given to the document type in 'Announcement' folder followed by 'Test/Quiz' folder as the 2nd priority, 'Lecture Notes' and 'Others' folder will the 3rd and 4th priority.

Once the priority assignment, and sorting the documents, then it will be broadcasted to the Student agent to inform the availability of these newly uploaded documents in the E-learning. This prioritize protocol is solely for the internal agent in E-Learning system since only the ElearningRequestor and ElearningExecutor agent are allowed to execute this protocol. Fig. 7 gives the overview of the possible states that exists during the prioritizing.

When the verification is launched, however, only one formulae is declared as 'True' whereas the remaining showing the negative result 'False' (refer Fig. 8). Fig. 9 and Fig. 10 shows the results showing the directed graph that is able to reach the specified state and the other graph are not able to move to the next state at all.

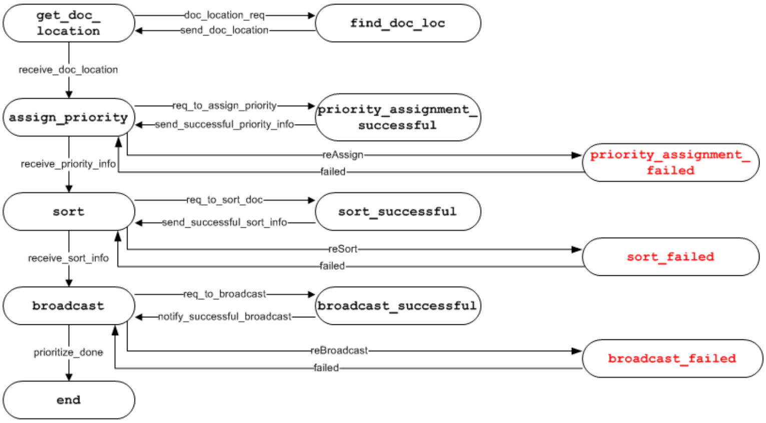
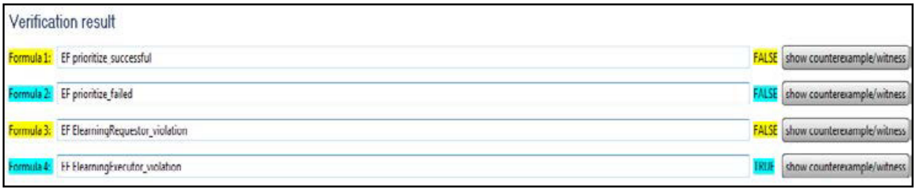


Fig. 7. State Diagram For Prioritize Protocol

Broadcast Protocol. Broadcast protocol is used to inform Student agent the availability of new documents uploaded by Lecturer. Fig. 11 shows the states diagram for both communicating agents. This diagram then is converted into ISPL file in in order to test each state for broadcast protocol. Fig. 12 and Fig. 13, on the other hand, show the verification result generated by MCMAS indicating a positive result for each expression evaluated.



Verification result

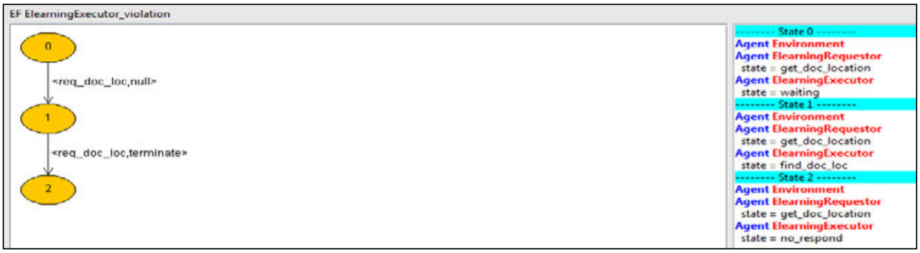
Formula 1: EF prioritize successful FALSE show counterexample/witness

Formula 2: EF prioritize_failed FALSE show counterexample/witness

Formula 3: EF ElearningRequestor_violation FALSE show counterexample/witness

Formula 4: EF ElearningExecutor_violation TRUE show counterexample/witness

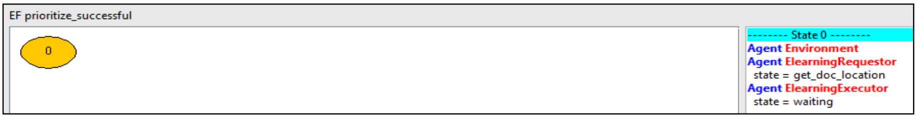
Fig. 8. Verification Result For Prioritize Protocol



EF ElearningExecutor_violation

```
-------- State 0 --------
Agent Environment
Agent ElearningRequestor
  state = get_doc_location
Agent ElearningExecutor
  state = waiting
-------- State 1 --------
Agent Environment
Agent ElearningRequestor
  state = get_doc_location
Agent ElearningExecutor
  state = find_doc_loc
-------- State 2 --------
Agent Environment
Agent ElearningRequestor
  state = get_doc_location
Agent ElearningExecutor
  state = no_respond
```

Fig. 9. All States Are Reachable In Formula 4 (Prioritize Protocol) - TRUE



EF prioritize_successful

```
-------- State 0 --------
Agent Environment
Agent ElearningRequestor
  state = get_doc_location
Agent ElearningExecutor
  state = waiting
```

Fig. 10. All States Are NOT Reachable In Formula 1 (Prioritize Protocol) - FALSE

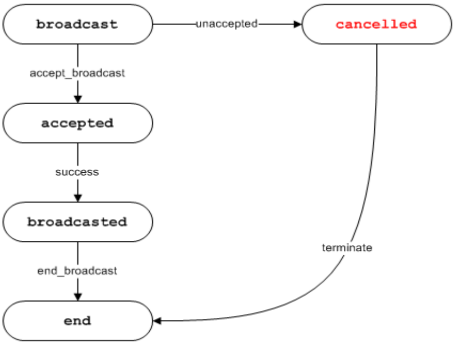
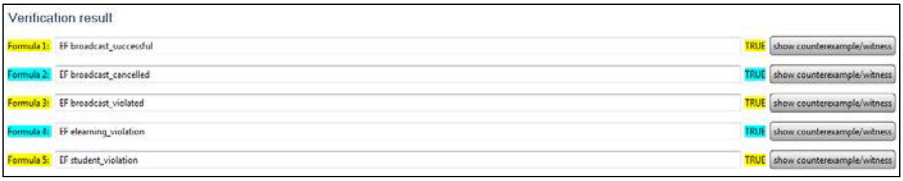


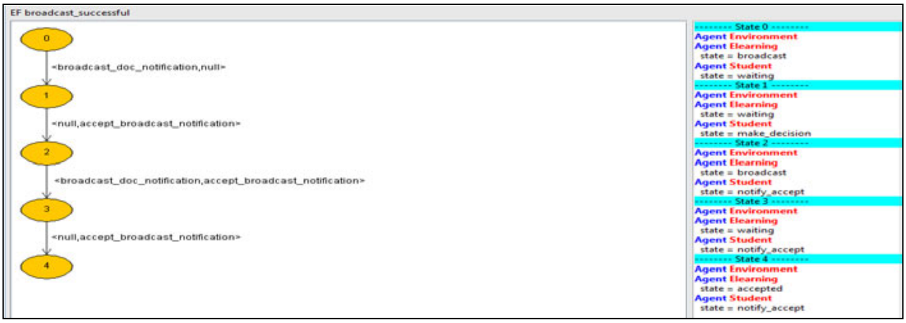
Fig. 11. State Diagram For Broadcast Protocol



Verification result

Formula 1:	EF broadcast_successful	TRUE	show counterexample/witness
Formula 2:	EF broadcast_cancelled	TRUE	show counterexample/witness
Formula 3:	EF broadcast_violated	TRUE	show counterexample/witness
Formula 4:	EF elearning_violation	TRUE	show counterexample/witness
Formula 5:	EF student_violation	TRUE	show counterexample/witness

Fig. 12. Verification Result For Broadcast Protocol



EF broadcast_successful

```
-------- State 0 --------
Agent Environment
Agent Elearning
 state = broadcast
Agent Student
 state = waiting
-------- State 1 --------
Agent Environment
Agent Elearning
 state = waiting
Agent Student
 state = make_decision
-------- State 2 --------
Agent Environment
Agent Elearning
 state = broadcast
Agent Student
 state = notify_accept
-------- State 3 --------
Agent Environment
Agent Elearning
 state = waiting
Agent Student
 state = notify_accept
-------- State 4 --------
Agent Environment
Agent Elearning
 state = accepted
Agent Student
 state = notify_accept
```

Fig. 13. All States Are Reachable In Formula 1

Download Protocol. Download protocol is used to download the broadcasted document and this interaction is between Student agent and E-Learning agent. The common process for downloading a document is shown in Fig. 14. From the verification result in Fig. 15, all formulae tested are showing a positive result which is giving 'True' which means all states prescribed are reachable. The directed graph in Fig. 16 proves the reachable property.

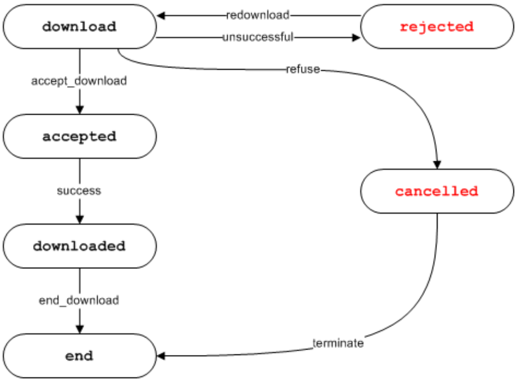
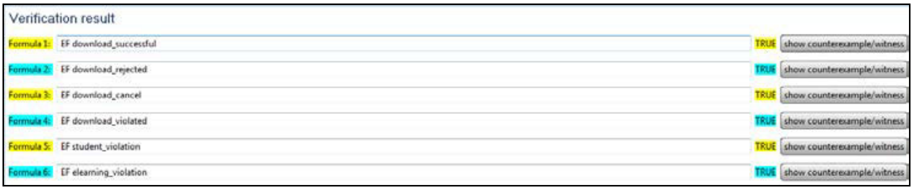


Fig. 14. State Diagram for Download Protocol



Verification result

Formula 1: EF download_successful TRUE

Formula 2: EF download_rejected TRUE

Formula 3: EF download_cancel TRUE

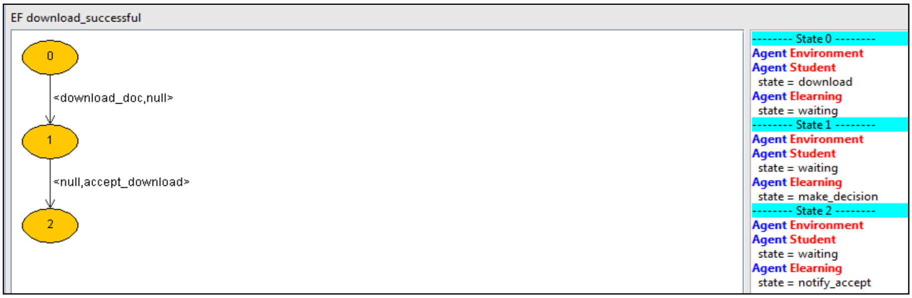
Formula 4: EF download_violated TRUE

Formula 5: EF student_violation TRUE

Formula 6: EF elearning_violation TRUE

show counterexample/witness
show counterexample/witness
show counterexample/witness
show counterexample/witness
show counterexample/witness
show counterexample/witness

Fig. 15. Verification Result For Download Protocol



EF download_successful

```
-------- State 0 --------
Agent Environment
Agent Student
 state = download
Agent Elearning
 state = waiting
-------- State 1 --------
Agent Environment
Agent Student
 state = waiting
Agent Elearning
 state = make_decision
-------- State 2 --------
Agent Environment
Agent Student
 state = waiting
Agent Elearning
 state = notify_accept
```

Fig. 16. All States Are Reachable In Formula 1

5 Conclusion

The result derived from MCMAS illustrates that three out of four protocols are showing a positive result which is giving ‘*True*’ answer for each formula evaluated except in prioritize protocol.

A further improvement on prioritize protocol need to be checked carefully in order to achieve a positive result. By achieving a positive result in each protocol, this study then can be considered as a successful and complete protocol since by using model checking approach, an exhaustive search of all possible states are identified. This is true based on the extended version of state diagram created for each protocol.

Acknowledgments. A token of appreciation to finish this paper goes to Mr. Fadzil Hassan as the supervisor who without fail assisting the author from the beginning to understand the theoretical aspect of this research. Another gratitude is also goes to Mr. Hilmi Hasan, as the co-supervisor who sometimes support the author in terms of technical aspect. Not to forget, a special thanks to MCMAS technical support team, Dr. Franco Raimondi, Dr. Hongyang Qu and Dr. Alessio Lomuscio from Imperial College London for assisting the author to troubleshoot the MCMAS toolkit.

References

1. Kannan, Kasmuri.: Agent Based Content Management for E-Learning (2005)
2. Gascuena, Fernandez-Caballero.: An Agent-Based Intelligent Tutoring System for Enhancing E-Learning/E-Teaching (2005)

3. Raimondi, F.: Model Checking Multi-agent Systems. Department of Computer Science. University of London, London (2006)
4. Edmund, J., Clarke, M., et al.: Model Checking, p. 307. The MIT Press, Cambridge (1999)
5. Latif, N.A., et al.: Utilizing Electronic Institution for protocol specification in agent-based e-learning system. In: Proc. of IEEE Student Conference on Research and Development (SCOReD), pp. 132–135 (2009)
6. Berard, B., et al.: Systems and Software Verification - Model-Checking Techniques and Tools, p. 196 (2001)
7. Huth, M.R.A., Ryan, M.: Logic in computer science: modelling and reasoning about systems, p. 387. Cambridge University Press, Cambridge (2000)
8. Raimondi, F.: Model Checking CTL - A gentle introduction (2007-2008)
9. Wooldrige, M.: An Introduction to MultiAgent Systems. John Wiley and Son, Chichester (2009)
10. Raimondi, F.: Computational Tree Logic And Model Checking - A Simple Introduction (2007-2008)
11. Lomuscio, A., Qu, H., Raimondi, F.: MCMAS: A model checker for the verification of multi-agent systems. In: Bouajjani, A., Maler, O. (eds.) CAV 2009. LNCS, vol. 5643, pp. 682–688. Springer, Heidelberg (2009)
12. Raimondi, F.: MCMAS Model Checking Multi-Agent Systems (2007-2008)

A New Stream Cipher Using Natural Numbers Based on Conventional Encryption Techniques: MINNSC

Rajendra Hegadi

Department of Computer Science and Engineering,
Pragati College of Engineering and Management
Raipur, C.G., India
rajendra.hegadi@gmail.com

Abstract. Stream cipher is as an important part of symmetric crypto system. One-time-pad cipher is the basic idea for stream ciphers, which uses XOR operation on the plain text and the key to generate the cipher. This research proposes a new stream cipher called MINNSC, based on the encryption decryption process using natural numbers. For a chosen natural, number blocks of non-zero divisors are generated and used as the encryption keys. These non-zero divisors form a group, their distribution is quite a random in nature, and this randomness is the desired property of stream cipher.

Keywords: Encryption, Decryption, Stream Cipher, Natural Numbers, Non-zero divisors.

1 Introduction

Randomness is the desired property of the symmetric encryption and decryption crypto system. In the proposed system we chose a natural number (say *k*) and generate a set of non-zero divisors which forms a group under modulo *k* [4]. The distribution of these non-zero divisors and their multiplicative inverses are quite a random in nature for different chosen natural numbers. We exploit this property of randomness to develop proposed MINNSC stream cipher.

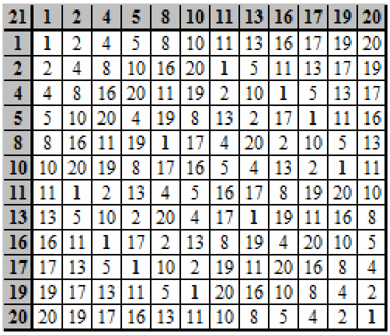
2 MINNSC- Proposed Stream Cipher

2.1 Multiplicative Group of Non-zero Divisors

The set of all the integers relatively prime to integer *k*, form a multiplicative group modulo *k*. That is, let *k* be any positive integer, then $G_k = \{x \in N \mid x \neq 0, (x, k) = 1\}$, where (x, k) is GCD and G_k is the set of all integers relatively prime to *k*. This set forms a multiplicative group under mod *k*, and the elements of this group are called as non-zero divisors. The Euler totient [1, 3] function $\varphi(k)$ gives the number of non-zero divisors i.e. number of relatively prime to *k* in the group G_k.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 427–436, 2011.

Table 1. Multiplication table for k = 21 with its relatively prime integers



21	1	2	4	5	8	10	11	13	16	17	19	20
1	1	2	4	5	8	10	11	13	16	17	19	20
2	2	4	8	10	16	20	1	5	11	13	17	19
4	4	8	16	20	11	19	2	10	1	5	13	17
5	5	10	20	4	19	8	13	2	17	1	11	16
8	8	16	11	19	1	17	4	20	2	10	5	13
10	10	20	19	8	17	16	5	4	13	2	1	11
11	11	1	2	13	4	5	16	17	8	19	20	10
13	13	5	10	2	20	4	17	1	19	11	16	8
16	16	11	1	17	2	13	8	19	4	20	10	5
17	17	13	5	1	10	2	19	11	20	16	8	4
19	19	17	13	11	5	1	20	16	10	8	4	2
20	20	19	17	16	13	11	10	8	5	4	2	1

Example: Table 1 gives set of all integers that are relatively prime to k = 21. This set G_{21}= { 1,2,4,5,8,10,11,13,16,17,19,20} can be easily proved as multiplicative group under mod 21. Many algorithms are proposed on encryption and decryption process using Latin Square or quasi groups [9]. Unlike Latin square, our proposed crypto system works on groups of non-zero divisors in modulo arithmetic.

2.2 MINNSC Algorithm

The proposed MINNSC algorithm has following steps:

i. Selecting a suitable integer k.

ii. Populating vector N with elements of group of non-zero divisors and vector IN with corresponding inverses of elements of N.
Example: For k=21 from table 1. N= {1,2,4,5,8,10,11,13,16,17,19,20} and
IN={1,11,16,17,8,19,2,13,4,5,10,20}

iii. Generation of key stream, the blocks of non-zero divisors from vector N.

iv. Key generator part: is responsible to generate a key stream independently from cipher and plain text.

v. Ciphering plain text.

The proposed MINNSC algorithm and the steps involved are explained below.

Algorithm: MINNSC

k is used as the key (let us call it main key) to populate two vectors N and IN.

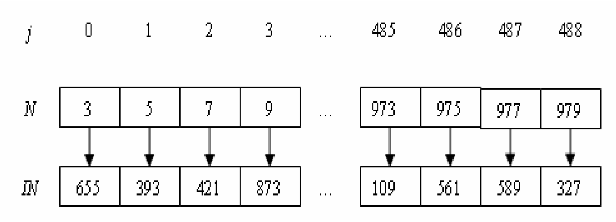
i. Vector N with non-zero divisors, the elements of this vector form a group under mod k

ii. Vector IN with inverses of each of non-zero divisor in vector N is also a group mod k

The key k in this algorithm is of variable length, sub keys for the encryption of stream of plain text data are generated from the main key k. The following are the steps involved in the encryption algorithm.

Step 1: Initialization of vector *N* and vector *IN*

The vector *N* is initialized with the group of non-zero divisor elements modulo *k*. The vector *IN* is initialized with the respective inverses of elements of vector *N*. Fig. 1 illustrates the initialization of these vectors for a natural number *k*= 982, excluding 1 and *k*-1=981 (for all *k* >1) because the inverses of these numbers are themselves. The upper bound of vector *N* and *IN* is 489 < *k*. The mapping of inverse elements is reversible and the extended Euclid's function [3] is used to compute the inverse elements.



j	0	1	2	3	…	485	486	487	488
N	3	5	7	9	…	973	975	977	979
IN	655	393	421	873	…	109	561	589	327

Fig. 1. Initialisation of vectors *N* and *IN* for key value *k*= 982

Step 2: Key stream generation

In this step, the vector *N* is divided in to pseudorandom blocks of size 26 each to produce another vector *B* which is the key stream. For key value *k* = 982 the number of non-zero divisors are 489, and hence the number of pseudorandom blocks of vector *B* are 19 (B0 to B18) with last block having elements fewer than 26. Each number in the every pseudorandom block vector *B* is assigned with the alphabets A to Z in sequence and similarly in vector *IN*. Similarly vector *IN* is also divided in to blocks of size 26 each in sequence.

To induce the Shannon confusion and diffusion and thwart the cryptanalysis [2], these pseudorandom blocks are created taking the vector elements randomly. However the mapping of elements and their inverses is still intact as the mapping is on-to-one. This on-to-one mapping keeps encryption and decryption reversible.

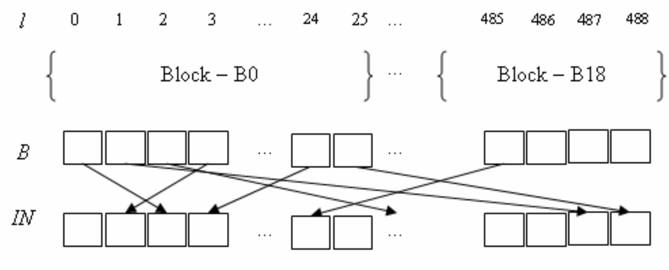


Fig. 2. Initialisation of pseudorandom blocks

Step 3: Stream ciphering

Algorithm: Encryption

Fig. 3 describes the generation of stream cipher using MINNSC. The plain text is stored in vector *S,* out put vector *C* contains the cipher text.
Input: vector *S*[*i*] – input text stream, where $0\leq i <$ number characters in plain text
Output: vector *C*[*j*]- output cipher stream, where $0\leq j <$ number characters in plain text.

For each *S*[*i*] assign the corresponding *B*[*j*] as per the 26 characters assignment and then swap the corresponding inverse element from vector *IN*[*l*], the respective alphabet for the *IN* vector block is the cipher text *C*. The process is continued till the end of stream vector *S*.

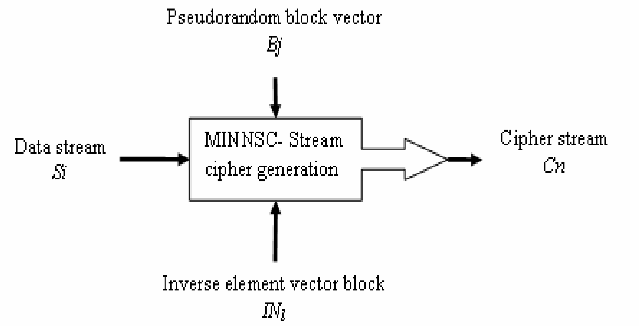

Pseudorandom block vector
Bj
Data stream
Si
MINNSC- Stream cipher generation
Cipher stream
Cn
Inverse element vector block
INl

Fig. 3. Encryption of stream of data - MINNSC

The function *Encrypt* accepts a character *S*[*i*] from stream vector *S* and returns the cipher character *C*[*n*] to the vector *C*. It uses function *binarysearch*() to search the inverse element from vector *IN* for a given number from vector *N*.

```
/* Stream Cipher Generation */
function_ Encrypt (S[i])
{         Num =0, InvNum = 0, Ch =‵ ′, RandNum = 0;
          Temp_ch  =S[i];
          Cipher_Ch   =     ‵ ′;
/* initialize ch to the plait text character */
          ch =S[i];
/* Now generate a random number from 0 to the max number of blocks in the pseudorandom blocks */
          RandNum = function_RandNumerGen();
          Num =function_returnNum ( $B_{RandNum}$ [j], ch )
/* Now assign the corresponding inverse number from
vector IN[1] to InvNum using binary search algorithm */
          InvNum = function_binsearch (B[j]);
          Cipher_Ch =  function_returnChar(InvNum);
          return (Cipher_Ch)            }
/* The function function_returnNum returns the number
corresponding to stream character assigned to a element
in the RandNum^th block in vector B, if Temp_Ch = "A"
then Num = first number in RandNum^th block and so on..*/
function_returnNum ( BlockNum, ch  )
{         return NonZeroDivisor_from_selectedBlock; }
/* function_returnChar returns a character from a block
in vector IN corresponding to the InvNum number */
function_returnChar(InvNum)
{
return Character_from_corresponding_block_in_vector_IN;
}
```

The keys streams *K1*, *K2*… are the pseudorandom blocks from vector *B*. The stream of plain text can be one character or a set of characters. To induce more randomness the plain text stream can be divided in to blocks of variable length and then appended with zero divisors to make them of equal size.

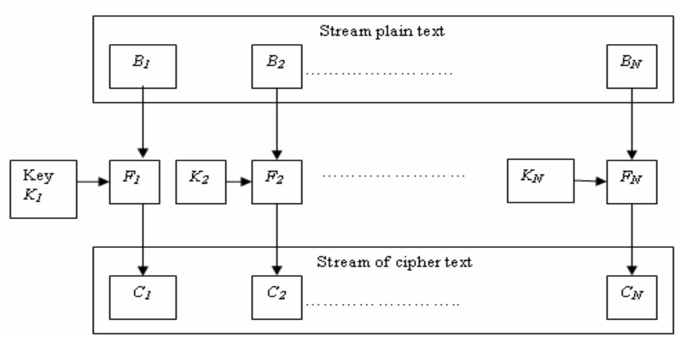


Fig. 4. MINNSC – Encryption block diagram

Fig. 5 gives the decryption block diagram which is the reverse of encryption process as the proposed cryptosystem is developed on modulo arithmetic system.

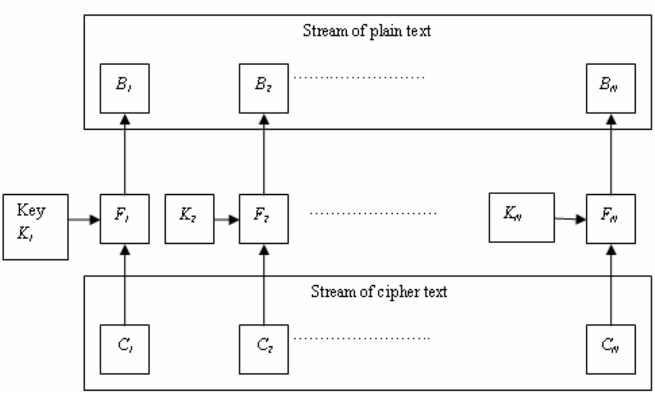


Fig. 5. MINNSC – Decryption block diagram

3 MINNSC with Feedback System for Avalanche Effect

Avalanche effect is also a desirable property of block ciphers because each output depends on the all the input. It states that an average of one half of the out put cipher text should change whenever there is a change in the input plain text. However for the non-feed back stream ciphers this property does not apply, as every out put stream cipher text is just encryption of the plain text stream text using encryption function with input as plain text stream and key stream only.

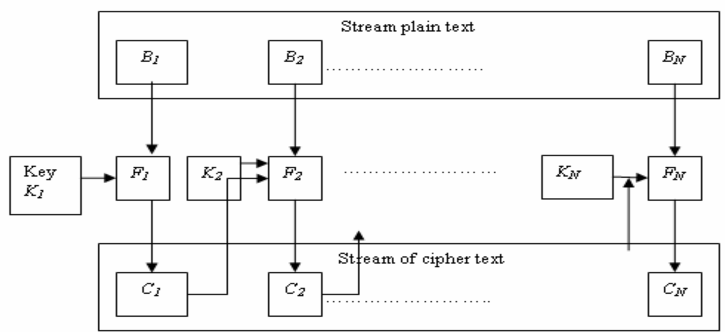


Fig. 6. MINNSC stream cipher encryption with feed back system

The MINNSC stream cipher with feedback system can achieve avalanche effect as shown in Fig.6. Here encryption function has three inputs, plain text stream, key stream and cipher text stream. The input cipher text is the out put of previous stage encryption and is the feedback for the current stage. The decryption function too has three inputs cipher text stream, key stream and the plain text stream of the previous stage and again essentially decryption with feed back system is reverse of that of encryption function.

But this feedback system is prone to related key attack on the cipher text, as each input to encryption function is dependent on the previous cipher text. There is an amount of predictability due to the same reason; hence it compromises the randomness which is the strength of non-feedback system.

4 Cryptanalysis of MINNSC

This section describes the possible attacks on the proposed MINNSC stream cipher and consequences. In this research, the most common (below discussed) methods are applied for the cryptanalysis of MINNSC.

4.1 Cipher Text Only Attack

This is an attack against cipher text when only the cipher text itself is available. Given only some information about *m* cipher texts, the attack has to have some chance of producing some information about the plain texts. As there is no linear or any other kind of mathematical relationship between the *m* different cipher texts, cipher text only attack is not effective. However the set of pseudorandom block keys are used for set of plain text for ciphering, there is possibility of deriving the natural number by working on all possible permutations of pseudorandom block keys.

4.2 Known Plain Text Attack

In this cryptanalysis the attacker knows or can guess the plaintext for some parts of the cipher text. The task is to decrypt the rest of the cipher text blocks using this information, for example frequency analysis. However the frequency analysis of the cipher text (discussed in next section) reveals very feeble information to the hacker to crack the cipher text or key.

4.3 Chosen Plain Text Attack

Here the attacker is able to have any text encrypted with the unknown key or guessing a key. The task in the chosen plaintext attack is to determine the key used for encryption.

This research found that proposed MINNSC stream cipher is highly resistive to the cipher text only, known plain text, chosen Plaintext attacks.

4.4 Differential Cryptanalysis

Differential cryptanalysis is effective for the attacks on block ciphers. The design of the stream cipher is such that, the difference between two cipher text reveal nothing about the secrete key. The MINNSC stream cipher does not offer a proper support for initial vectors but allows variable key length. The solution for the above security problem can be implemented with choosing several initial vectors with particular difference between them. Proper implementation of randomness can deceive the differential cryptanalysis. MINNSC stream cipher offers several levels of randomness, choosing initialization vector, selecting the random block of key from blocks of non-zero divisors, selecting the variable length of plain text, repeated application of the MINNSC and a feed back system.

4.5 Related Key Cryptanalysis

It is similar to differential cryptanalysis, but it examines the differences between keys. In this attack a relationship is chosen between a pair of keys, but does not know the keys themselves. It relies on simple relationship between sub keys in adjacent rounds, encryption of plain texts under both the original (unknown) key *k*, and some derived keys K_1,K_2….. It is needed to specify how the keys are to be changed; there may be flipping of bits in the key without knowing the key.

There are two classes of related-key attacks [8]. First attack uses related key plaintext pairs for which most of the encryption function is equivalent. Such relations exist when the key schedule is very simple. Second attack is that treat the key relation as another freedom level in the examination of statistical properties of the cipher for example weak relative round functions. On the other hand, once such a relation can be found, it can be used to devise an attack on the cipher.

MINNSC stream cipher allows generating multiple non related-keys by choosing random blocks of non-zero divisors. Hence the key schedule of MINNSC is a strong one in turn leading to strong opposition on attacks on encryption functions.

5 Analysis of MINNSC

This section gives the analysis of the proposed stream cipher MINNSC.

5.1 Frequency Analysis of MINNSC

The Frequency analysis of proposed stream cipher is divided into two parts as explained below [3].

5.1.1 Frequency Analysis of Cipher Text

In this test number of occurrence of each character in cipher text is counted, compared and replaced with the equivalent standard relative English alphabet, resulting a reasonable skeleton of the message. A more systematic approach is also followed like, certain words known to be in the text or repeating sequence of cipher letters and try to deduce their plain text equivalents. The process is repeated to digraphs/ digrams and trigraphs/ trigrams. The entire process is repeated with changing the number of characters per blocks of plain text. For the test to be more effective, the plaintext and hence the cipher text length is chosen more than 40000 characters.

From the results it is evident that the frequency analysis of cipher text does not reveal much to the hacker as the percentage of match of characters is very low and increase in the number of plain text stream characters per block also does not change the percentage of match drastically and hence their exists a nonlinearity.

5.1.2 Relative Frequency Analysis of Cipher Text

It is another way of revealing the effectiveness of the algorithm. The number of occurrences of each character in the text is counted and divided by the number of occurrences of the letter *e*, the most frequently used letter, to normalize the plot. As the result, *e* has a relative frequency of 1, *t* of about 0.76 and so on for the plain text. The points on the x- axis correspond to the letters in order of decreasing frequency.

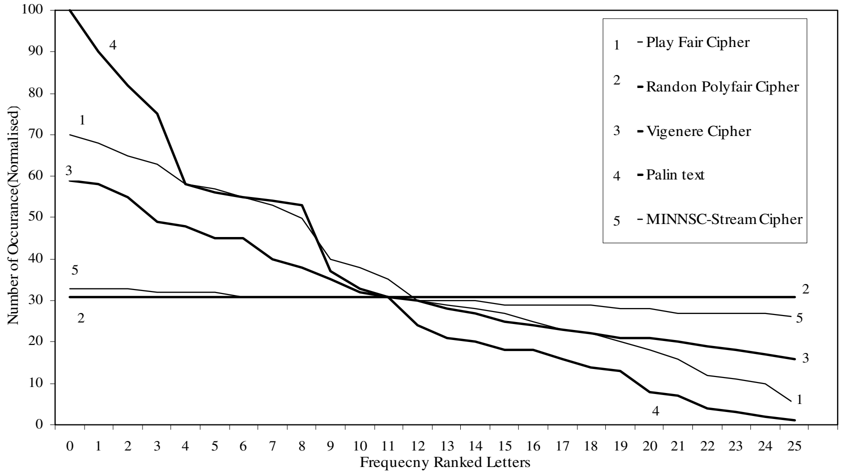


Fig. 7. Relative frequency analysis of cipher text

Fig. 7. therefore shows the extent to which the frequency distribution of characters, which makes it trivial to solve substitution ciphers, is marked by encryption. The cipher text plot for the MINNSC stream cipher is much flatter than that of the other cipher algorithms, and hence the cryptanalysis using cipher text only is almost impossible.

5.2 Exhaustive Key Search for MINNSC

The exhaustive key search is a brute-force search. It is the basic technique of trying every possible key in turn until the correct key is identified. The proposed stream cipher has possible exhaustive key search attack. Assuming a fast computer applied to the problem of solving a cryptogram by this trial-and-error procedure. The computer would cycle through the possible permutations of each block 26 invertible elements, checking if the result were reasonable.

Let *k* be the natural number used for the MINNSC stream cipher and let *m* be the number of non-zero divisors in the group deduced from *k*. The Euler totient function $\varphi\ (k) = m$ gives us number of non-zero divisors and the average value of *m* is given by $3n/\pi^2$ [1].

For example for $k = 12345677$ then there are $m = 11919937$ possible non-zero divisors, now the keys for stream cipher are the blocks of non-zero divisors that are generated with each block having 26 elements corresponding to the 26 letters. Then the number permutations of blocks of non-zero divisors can be created are $kPr \Longrightarrow {}_KP_{26} = 9.62 \text{ X } 10^{183}$ numbers of different methods the pseudorandom blocks and hence the keys for encrypting the plain text can be created. This is an enormous number of possibilities.

Assuming that the computer could check 10,000 million permutations per second (100 GHz Speed - which is optimistic since there would be considerable effort to determine if the result were reasonable), it would take about – (9.62 X 10^{183}) / (2 X 10^{11}) = 4.8 ×10^{172} seconds =1.52 × 10^{165} Years are required to complete the computation. Hence the exhaustive key search or brute force analysis doesn't yield anything to the hacker.

6 Conclusion

The aim of the presented work is to introduce a new stream based on the conventional encryption technique substitution cipher. Desired property randomness is implemented at various stages. The ciphers that are generated by the proposed stream cipher method have been analyzed and discussed. All possible attacks on the presented algorithm are discussed; it is shown that the algorithm is very simple and easy to implement and equally hard to crack.

References

1. Tom, M.: Apostal: Introduction to Analytic Number Theory. Springer, Heidelberg (1976)
2. Koblitz, N.: A Course in Number Theory and Cryptography. Springer, New York (2002)
3. Stallings, W.: Cryptography and Network Security: Principles and Practices. PHI (2004)
4. Hegadi, R., Nagaraj, M., Patil, S.S.: Encryption and Decryption Process using Composite Numbers. International Journal of Computer Science and Network Security, IJCSNS 7(1), 371–377 (2007)
5. Robshaw, M.J.B.: Stream Ciphers. Technical Report TR-701, version 2.0, RSA Laboratories (1995)
6. Pal, J.K. Mandal, J. K., Gupta, S.: Composite Transposition Substitution Chaining Based Cipher Technique. ADCOM, pp433-439 (2008)
7. Zhang, Y.-P., Sun, J., Zhang, X.: A stream Cipher Algorithm based on conventional Encryption Techniques. In: CCECE, pp. 649–652 (2004)
8. Biham, E.: New Types of Cryptanalytic Att acks Using Related Keys. Journal of Cryptology 7(4), 229–246 (1994)
9. Pal, S.K., Sumitra: Development of Efficient Algorithms for Quasigroup Generation & Encryption. In: IEEE International Advance Computing Conference-IACC, pp. 940–945 (2009)

Remote User Authentication Scheme with Hardware-Based Attestation

Fazli Bin Mat Nor[1], Kamarularifin Abd Jalil[1], and Jamalul-lail Ab Manan[2]

[1] Faculty of Computer & Mathematical Sciences, Universiti Teknologi Mara,
40450 Shah Alam, Selangor, Malaysia
[2] MIMOS Berhad, Technology Park Malaysia, 57000 Bukit Jalil,
Kuala Lumpur, Malaysia
fazlimnor@hotmail.com, kamarul@tmsk.uitm.edu.my,
jamalul.lail@mimos.my

Abstract. Many previous works on remote user authentication schemes are related to remote services environment such as online banking and electronic commerce. However, these schemes are dependent solely on one parameter, namely, user legitimacy in order to fulfill the authentication process. Furthermore, most of the schemes rely on prearranged shared secret key or server secret key to generate session key in order to secure its communication. Consequently, these schemes are vulnerable to malicious software attacks that could compromise the integrity of the platform used for the communication. As a result, user identity or shared secret key potentially can be exposed due to limitation of the scheme in providing trust or evidence of claimed platform identity. In this paper, we propose a remote authentication with hardware based attestation and secure key exchange protocol to resist malicious software attack. In addition, we also propose pseudonym identity enhancement in order to improve user identity privacy.

Keywords: remote user authentication; remote attestation; trusted platform module; privacy; pseudonym.

1 Introduction

In today's environment, web services such as online banking and e-commerce have become more important as more and more users are depending on these services to manage their daily life business. These services normally require user to be authenticated in order to gain access to the services. Hence, remote authentication becomes an important security measurement to verify legitimacy of the user. In 1981, Lamport [1] first introduced verifier-based remote user authentication over insecure communication between user and the services. Since then, many remote authentication schemes have been proposed to improve the scheme. However, user credential alone is not enough to secure the communication between user and the remote services because user's platform used in the communication is always vulnerable to any attacks such as malicious software attacks. Furthermore, some of past works on remote authentication schemes [10,11,12,13] use prearranged shared secret key or server secret key to

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 437–447, 2011.

generate session key in order to secure communication between client and server. Consequently, if the system is compromised, user credential or secret keys can potentially be stolen. Therefore, additional security measures such as platform integrity verification is needed to detect any illegitimate changes to platform configuration as well as providing secure secret key exchange.

Apart from validating user legitimacy, we certainly need another mechanism to check platform integrity, which also becomes an important criterion in authentication process especially if we need to detect any changes to platform configuration caused by malicious software or malware. For instance, if any of the communicating parties has been infected by malicious software, malicious payload integrated in the malware normally opens a pathway to create dangerous activities such as stealing user or server secret keys from its victim and later takes control of the system without owner's knowledge. Thus, without platform integrity verification, any prearranged shared secret keys used or kept by communicated parties are vulnerable to malware attacks and potentially exposed if it is not securely protected.

In order to effectively perform platform integrity checks mentioned above, Trusted Platform Module (TPM) [18] based remote attestation has been chosen because of its capability to provide attestation based information about the platform and to ensure integrity of the platform is not tampered. Furthermore, communication between client and server would be more secure if secret key used in the communication is not stored anywhere or does not require to be sent across the network. For this reason, we have chosen Secure Remote Password (SRP) [2] protocol to fulfill secure secret key exchange requirements.

1.1 Secure Remote Password (SRP) Protocol

Secure Remote Password (SRP) protocol is password authentication and key exchange protocol over an untrusted network and it has been developed based on zero knowledge proof and verifier based mechanism [2]. In the event of authentication, zero knowledge proof allows one party to prove themselves to another without revealing any authentication information such as password. On the other hand, verifier based mechanism requires only verifier that has been derived from password to be stored in the server side. Thus, this protocol makes sure no sensitive authentication information such as password to be sent across the network.

The SRP protocol as shown in Figure 1 consists of two stages. First stage of the protocol is to set up authentication information of the client and store the information on the server side. At this stage, client calculates secret information sent by the verifier based on client's password and random salt. Server then stores client's username (i), verifier (v) and random salt (s) for authentication purposes. Second stage is the authentication process. Steps of SRP authentication are follows [2]:

1. Client then generates a random number (a) and by using generator (g), client calculates public key (A) = g^a. Client starts the authentication process by sending public key, A with its username (i) to the server.
2. Server looks up for client's verifier (v) and salt (s) based on the username (i). Server then generates its random number (b) and computes its public key (B) using verifier (v) and generator (g) and server sends (s, B) to client.

3. Upon receiving B and s, client calculates private key (x) based on salt (s) and its password (p).
4. Both client and server then compute their own session key (S) with different calculation method. Session key (S) calculated by both parties will match when password used in the calculation is originally used to generate the verifier (v).
5. Both sides then generate cryptographically strong session key (k) by hashing session key (S).
6. In order for client to prove to the server that it has correct session key, it calculates M_1 and sends to the server. The server verifies the M_1 received from client by comparing with its own calculated M_1 values.
7. Server then sends M_2 to client as evidence that it has correct session key.
8. Finally, once client verifies M_2 is matches with its own calculated M_2 value, client is now authenticated and secured communication channel can be established.

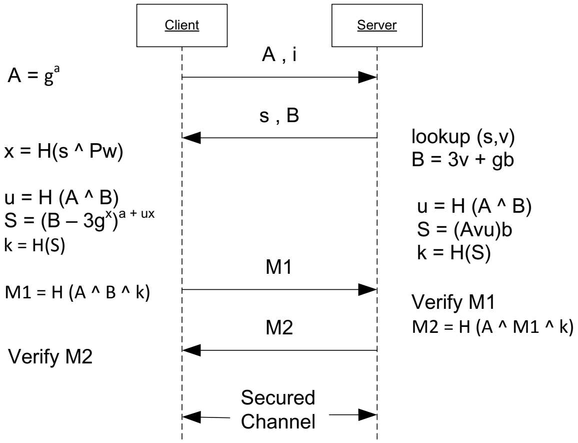


Fig. 1. Secure Remote Password authentication protocol

1.2 TPM Based Remote Attestation

Remote attestation allows remote host such as server to verify integrity of another host's (client) platform such as its hardware and software configuration over a network. Thus, by using this method, remote host will be able to prove and trust that client's platform integrity is unaffected by any malicious software. As mentioned by Trusted Computing Group (TCG) [14], an entity is trusted when it always behaves in the expected manner for the intended purpose. Therefore, remote attestation is an important activity to develop trust relationship between client and server to ensure the communication is protected from illegitimate entity.

In remote attestation, client's platform integrity is measured in relation to its hardware and application information and the integrity measurement values will be stored into a non-volatile memory in TPM. Next, the measurement values are integrated as part of integrity report that later will be sent to host system such as server to be analyzed and verified, in order to prove to the host system that its platform integrity is untouched by any unwanted entity.

However, TPM hardware itself is a passive chip. Therefore, TPM alone is unable to measure the platform and it requires software intervention to activate its functionalities. In order to recognize platform as trustworthy platform, platform measurement process has to start at boot time of its host. Trusted boot process such as TrustedGrub [15] measures platform components such as BIOS, boot loader and operating system and extends integrity measurement into 160 bit storage register inside TPM called Platform Configuration Register (PCR) [16, 17]. Therefore, this hardware based integrity measurements can be amalgamated with other application based measurements to produce evidence to other party in attestation process.

Nevertheless, integrity measurement alone cannot provide the identity of the platform. For this reason, each TPM has its unique Endorsement Key (EK) certified by its manufacturer which identifies the TPM identity. To overcome privacy concerns if EK is used directly in attestation process, Attestation Identity Key (AIK) which is derived from the EK is used to sign integrity measurement. Thus, TPM based remote attestation is also crucial to establish the truly trusted identity of the platform to other party.

1.3 Our Contribution

Trust and privacy are important security elements that must be taken care of when dealing with remote services. With this in mind, each parties involve in the communication must ensure that they communicate with legitimate and trusted entities as well as their identity privacy is protected. Thus, it is crucial to incorporate both these elements in authentication scheme related to remote services.

In this paper, we propose remote user authentication protocol that makes use of TPM features to incorporate trust element and protect user's privacy with pseudonym identity. In addition, we also take advantage of SRP key exchange protocol to provide strong session key in our protocol communication.

1.4 Outline

This paper is organized as follows: Section 2 discusses previous works on authentication related to remote services and their issues. Section 3 presents our proposed solution, whereas section 4 and 5 analyze security elements on the proposed protocol. Finally, section 6 concludes the paper.

2 Related Works

Many of past works on remote authentication protocol [10,11,12,13] have been proposed to overcome insecure communication between client and server. These protocols have solely focused only on user legitimacy and require shared secret key to provide secure communication. However, without any protection to endpoint platform at both client and server, these protocols are still vulnerable to malicious software attack when they reach the endpoints. Furthermore, user credential and shared secret key can be potentially exposed if it is not securely protected at endpoints.

Zhou et al. [4] took initiative to introduce password-based authenticated key exchange and TPM-based attestation in order to have secure communication channels and endpoint platform integrity verification, due to the issue of normal SSL/TLS or IPSec which do not provide endpoint integrity. They proposed Key Exchange with Integrity Attestation (KEIA) protocol which is based on a combination of both password-based authenticated key exchange and TPM-based attestation. This protocol is the first known effort that combines platform integrity to endpoint identity in order to prevent reply attack and collusion attack. KEIA adopts SPEKE [9] as their key exchange protocol. However, Hu [10] stated that SPEKE is susceptible to password guessing attack when simple password is used. On the other hand, KEIA protocol uses prearranged shared secrets as the part of their authentication.

Ali [6] has proposed remote attestation on top of normal SSL/TLS secure channel. His work provides architecture for access control based on the integrity status of the web client. Thus, client with compromised integrity will not be able to access services on the server. However, this solution relies solely on Secure Socket Layer (SSL) for their secure communication. Unfortunately, integrity reporting protocol cannot rely on SSL alone as the protocol is vulnerable to main-in-the-middle attack [4, 7]. Cheng et al. [7] proposed a security enhancement to the integrity reporting protocol by implementing cryptographic technique to protect measurement values. However, their solution requires client to generate premaster secret key and client has to carefully secure the key to avoid impersonation if the key is stolen [2].

3 Proposed Solution

In this section, we present remote user authentication protocol with both elements of trust and privacy. For this purpose we decided to use TPM based remote attestation and user identity pseudonymization as trust and privacy implementation method respectively. In addition, SRP is adopted in the proposed protocol as the secured key are being exchanged between communicating parties. Notations used in proposed protocol are describes in Table 1.

3.1 Protocol Description

Our proposed protocol as shown in Figure 2 consists of two phases; registration phase and verification phase. In registration phase, client sends pseudonym identity, verifier value, random salt value and public certificate of its AIK to the server to set up authentication information via secure channel. Following are the steps for registration process:

1. Client computes its pseudonym identity (u) by hashing combination of user identity and platform PCR values.
2. In order to compute private key (x), client generates random salt value (s) to be combined with client's password in hash function.
3. Client calculates its verifier value (v) derived from private key (x) using generator (g).

4. Client then sends u, v, s and public certificate of its AIK to server. Server then stores that information in database for authentication purposes.

Table 1. Notation of the proposed protocol.

Notation	Description
i	User identity (user name)
PCR	Selected PCR value
u	Client pseudonym identification
g	A primitive root modulo *n* (often called a generator). While *n* is a large prime number.
s	A random string used as the user's salt
Pw	The user's password
x	A private key derived from the password and salt
v	Password verifier
a,b	Ephemeral private keys, generated randomly and not publicly revealed, $1 < a$ or $b < n$
A,B	Corresponding public keys
H(.)	One-way hash function
m ^ n	The two quantities (strings) m and n concatenated
k	Session key
Mc	Client evidence
Ms	Server evidence
SML_c	Client's Store Measurement Log
SML_s	Known good hashes of Store Measurement Log (Server side)
Sn	Signature value signed with AIK private key
Enc_k	Encryption method with *k* as key
Dec_k	Decryption method with *k* as key

In authentication phase, there are two stages of proof evidence that need to be fulfilled in order to complete the authentication process. First stage, client and server need to proof each other that they are having the same session key (k) without revealing any information about the key. This is done based on zero-knowledge proof calculation implemented in SRP protocol. Second stage, client needs to provide proof to the server that its platform integrity is unaffected by any malicious software. Following are steps for authentication process:

1. Client calculates its pseudonym identity (u) by hashing combination of user identity (i) and selected platform PCR values. Client then calculates its public key (A) using generator (g) and sends both values (A,u) to server.
2. Server looks up client's salt value (s), verifier (v) and public AIK certificate from its database based on pseudonym identity (u) given by client. At the same time, server calculates its public key (B) using generator (g) and sends both values (s, B) to client. Prior to sending the values, server computes its session key (k) based on mathematical calculation stated in SRP protocol

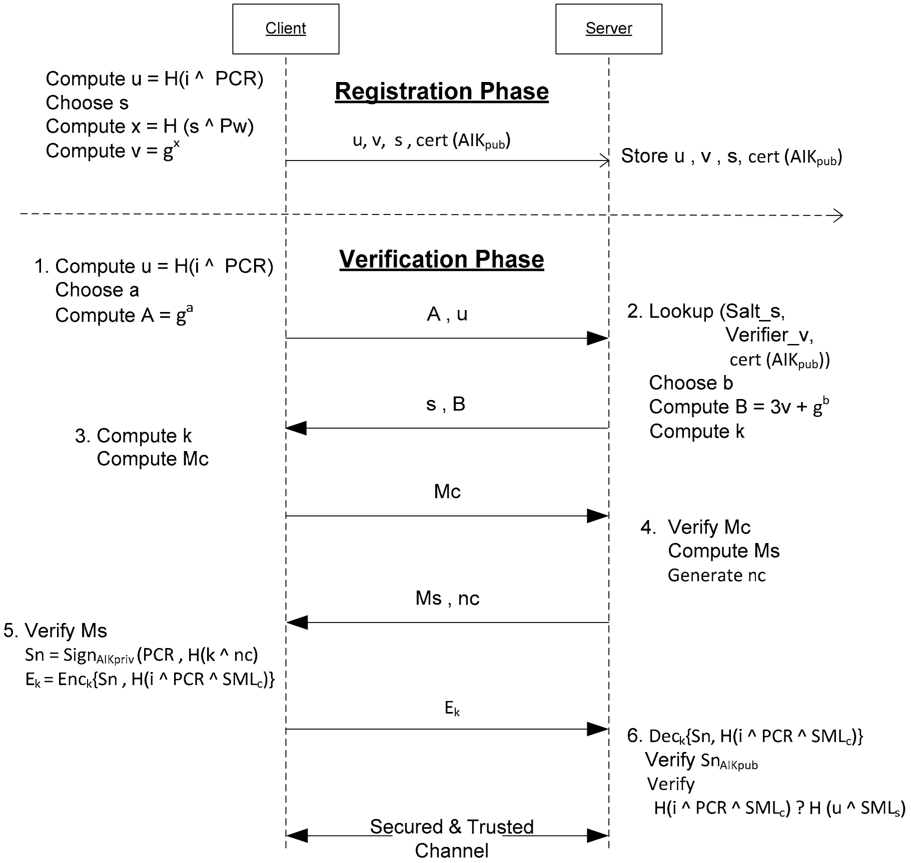


Fig. 2. Proposed solution registration and authentication scheme

3. Upon receiving salt (s) and server's public key (B), client computes its session key (k). Client then sends Mc as evidence that it has the correct session key.
4. Once the server verifies Mc is matched with its own calculated Mc, server then computes Ms to prove that it also has the correct session key (k). Server then sends Ms together with random number (nc) to client.
5. Client verifies Ms with its own calculated Ms, if the values matched, client then invokes TPM functionality by signing its platform measurement values stored in PCR with AIK private key. The signature is then encrypted with session key (k) together with hashed values of client's username (i), PCR values and stored measurement log (SMLc). Next, the encrypted values (Ek) are sent to the server.
6. Upon receiving Ek from client, server decrypts the Ek using its session key (k). Server then verifies signature (Sn) with client's AIK certificate. Once the signature is verified, server computes hashed values of pseudonym id (u) with its own stored measurement log (SMLs). Next, server verifies its

measurement hashed values with client’s measurement hash values. If the verification succeeds, both parties now able to communicate in secured and trusted channel.

4 Security Analysis

In this section, we analyze security elements on our proposed solution based on integrity verification, impersonation attack, stolen verifier attack, insider attack and identity protection.

4.1 Integrity Verification

One of the important security elements in our proposed solution is trust. In order to establish trust, client’s platform integrity needs to be analyzed by remote host. Therefore, it is crucial to secure client’s platform integrity measurement from any illegitimate parties. Our protocol assures the integrity measurement is transferred securely. This is done by encrypting the measurement with session key (k). Thus, in order to manipulate the integrity measurement value, attacker would need to capture (k), however it is impossible as (k) has never been exchanged between client and server. Furthermore, our solution uses TPM as tamper-proof hardware that protects all the measurements from being manipulated at client side.

4.2 Impersonation Attack

Attacker would not able to impersonate either client or server without knowing session key (k) as implementation of SRP protocol requires zero knowledge proof. Without session key (k), attacker would not able to compute evidence Mc or Ms, in order to prove he or she has the correct session key. Moreover, session key (k) is never passed over the network and this will make impersonation attack almost impossible.

4.3 Stolen Verifier Attack

Normally when password related information such as verifier is stored at the server side, it is vulnerable to stolen verifier attack. This attack happens when attacker able to gain access to the server and manage to extract verifier information from its database. This attack also might lead to impersonation attack when attacker manages to manipulate authentication process using the stolen verifier. The strength of our protocol is that even though attacker manages to steal the verifier (v), the attacker would not able to continue with authentication process without client’s password as it requires expensive dictionary search to reveal it [2].

4.4 Insider Attack

Weak client’s password or server secret key stored in server side is vulnerable to any insider who has access to the server. Thus, in the event of this information is exposed, the insider able to impersonate either party. The strength of our proposed protocol is that it does not store any client's password or server secret key in the server side. Therefore, our scheme can prevent the insider from stealing sensitive authentication information.

4.5 Identity Protection

Current remote user authentication protocols [4,6,11,7] lack privacy protection as most of the protocols have no mechanism to protect that information from being linked back to actual user. Our protocol preserves user identity privacy by replacing user identity with pseudonym identity (u) which is a hashed value of user identity and selected platform PCR values. Pseudonym identity is important because in the event of server's database has been compromised; user identity privacy is still protected due to the fact that attacker cannot manipulate the pseudonym identity or link it back to actual user.

5 Discussion

In this section, we summarized our protocol and other related schemes based on security analysis. Table 2 shows comparison between our scheme and other schemes.

Table 2. Security analysis summary

Protocols	Security Analysis				
	IV	IA	SV	IT	IP
Our scheme	√	√	√	√	√
Zhou et al. [4]	√	√	√	Ø	X
Ali [6]	√	√	√	X	X
Cheng et al. [7]	√	√	√	X	na
Hu et al. [10]	X	√	√	√	√
Liao et al. [11]	X	√	√	X	X
Chien et al. [12]	X	√	√	X	√
Chai et al. [13]	X	√	√	X	√

* Notation:
IV – Integrity Verification
IA – Impersonation Attack
SV – Stolen Verifier Attack
IT – Insider Attack
IP – Identity Protection

√ – Satisfied
X – Not satisfied
Ø – Partially satisfied
na – Unrelated

6 Conclusion

In this paper, we have shown current remote user authentication schemes require some improvement in terms of providing protection from malicious software attack and preserving user identity privacy. We propose trusted and secure remote user authentication with privacy enhancement to user identity in order to fulfill limitation of current schemes. The proposed solution incorporates TPM based attestation and

SRP key exchange protocol to provide trusted and secure communication between client and server. In addition, the proposed protocol preserves user identity privacy by replacing actual user identity with pseudonym identity. We demonstrate security analysis on proposed protocol based on a few security criteria which shows that the proposed protocol resists any possible threats.

Acknowledgement

The authors would like to thank Universiti Teknologi MARA for funding this research under the Excellence Fund.

References

1. Lamport, L.: Password authentication with insecure communication. Communications of the ACM 24(11), 770–772 (1981)
2. Wu, T.: The Secure Remote Password protocol. In: Internet Society Network and Distributed Systems Security Symposium (NDSS), San Diego, pp. 97–111 (1998)
3. Juang, W.S., Wu, J.L.: Efficient user authentication and key agreement with user privacy protection. International Journal of Network Security 7, 120–129 (2008)
4. Zhou, L., Zhang, Z.: Trusted channels with password-based authentication and TPM-based attestation. In: International Conference on Communications and Mobile Computing, pp. 223–227 (2010)
5. Zhang, M.: Analysis of the SPEKE password-authenticated key exchange protocol. IEEE Communications Letters 8(1), 63–65 (2004)
6. Ali, T.: Incorporating remote attestation for end-to-end protection in web communication paradigm. In: International Conference on Internet Technologies and Applications, Wales, UK (2009)
7. Cheng, S., Bing, L., Yang, X., Yixian, Y., Zhongxian, L., Han, Y.: A security-enhanced remote platform integrity attestation scheme. In: Wireless Communications, Networking and Mobile Computing (WiCom 2009), vol. 4, pp. 24–26 (2009)
8. Stumpf, F., Tafreschi, O., Röder, P., Eckert, C.: A robust integrity reporting protocol for remote attestation. In: Proceedings of the Workshop on Advances in Trusted Computing, WATC (2006)
9. Jablon, D.: Strong password-only authenticated key exchange. SIGCOMM Computing Communication 26(5) (1996)
10. Hu, L., Yang, Y., Niu, X.: Improved remote user authentication scheme preserving user anonymity. In: Communication Networks and Services Research (CNSR 2007), pp. 323–328 (2007)
11. Liao, Y.P., Wang, S.-S.: A secure dynamic ID based remote user authentication scheme for multi-server environment. Computer Standards & Interfaces 31(1), 24–29 (2009)
12. Chien, H.-Y., Chen, C.-H.: A remote authentication scheme preserving user anonymity. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA 2005), Washington, USA, vol. 2, pp. 245–248 (2005)
13. Chai, Z., Cao, Z.-F., Lu, R.: Efficient password-based authentication and key exchange scheme preserving user privacy. In: Cheng, X., Li, W., Znati, T. (eds.) WASA 2006. LNCS, vol. 4138, pp. 467–477. Springer, Heidelberg (2006)

14. Trusted Computing Group: TCG specification architecture overview, specification revision 1.4 (2007)
15. TrustedGrub, http://www.sirrix.com/content/pages/trustedgrub.htm
16. Kinney, S.: Trusted Platform Module Basics: Using TPM in Embedded System. NEWNES (2006)
17. Challener, D., Yoder, K., Catherman, R., Safford, D., Doorn, L.V.: A Practical Guide to Trusted Computing. IBM Press (2008)
18. Sadeghi, A.-R.: Trusted Computing — Special Aspects and Challenges. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M. (eds.) SOFSEM 2008. LNCS, vol. 4910, pp. 98–117. Springer, Heidelberg (2008)

Optimal Camera Placement for 3D Environment

Siti Kamaliah Mohd Yusoff[1], Abas Md Said[1], and Idris Ismail[2]

[1] Computer Information Sciences Department,
Universiti Teknologi PETRONAS, 31750 Malaysia
noradrelina@gmail.com, abass@petronas.com.my
[2] Electrical and Electronic Department,
Universiti Teknologi PETRONAS, 31750 Malaysia
idrisim@petronas.com.my

Abstract. Efficient camera placement is important in order to make sure the cost of a monitoring system is not higher than what it should be. This is also to ensure the maintenance of that system will not be complex and take longer time. Based on these issues, it has become an important requirement to optimize the number of the camera in camera placement system inside particular environment. This problem is based on the well-known Art Gallery Problem but most of previous works only proposed solution to this problem in 2D. We propose a method for finding the minimum number of cameras that can observe maximum space of 3D environment. In this method we assume that each of the cameras has limited field of view of 90o and only to be placed on the wall of the environment. Placement in 3D environment uses volume approach that takes frustum's volume and space's volume to calculate minimum number of camera.

Keywords: placement, optimal camera, sensor placement, visual network.

1 Introduction

Camera has been used largely in video surveillance and security system as a tool to observe and monitor a specific area and also for crime and hazard prevention. Inefficient camera placement in a surveillance system is one of the main reasons that increases cost and complexity of maintenance system. Sometimes design of the camera placement is not sufficient because fewer cameras are used. In other cases, the number of cameras used is more than enough. Any one of the situation may lead to maintenance problems in the future. By having the best placement for cameras, set-up cost is minimised and at the same time the maintenance cost can be reduced.

This paper deals with the problem of 3D space monitoring which is based on the famous 'Art Gallery Problem'. The goal in the art gallery problem is to find the minimum number of guards that can monitor a fixed number of paintings in a gallery. The layout of the art gallery is the polygon and the covering points (vertices on the polygon) are the guards. In this case the guards can be placed in the interior, on walls or on the corner of the gallery. Original art gallery theorem states that at most (n/3) guards are required for covering polygons with *n* edges [1]. Many variations of art gallery problem have been studied in previous works that address different

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 448–459, 2011.

specification for the guards and the polygons and also additional constraints (see for example [2, 3, 4, 5 and 6]). M. C. Couto *et al.*[2] focuses on the placement of the guards that is restricted to the vertices of the polygon with 360° FOV however C. D. Toth [6] addressed art gallery problem with guards of range of vision from 0° to 360° . D. Jang *et al.* [3] and S. K. Gosh [4] proposed algorithms for solving minimum vertex and edge guard problem for polygons with or without holes. In [5], the authors introduced a new concept that is the area of the art gallery to be observed is only limited to expensive paintings. For our research, the number of cameras is dynamic and the 3D space can be with or without holes.

Finding the optimal camera placement is a difficult problem because of the constraints that need to be taken into consideration. These are the complexity of the environment, diverse camera properties, and numerous number of performance metrics for different applications. Each solution is based on different requirements and constraints but has same objective that is to find the minimum number of cameras. We proposed a new method for camera placement in 3D environment because until now most of the proposed methods are only done in 2D workspace which is not applicable to 3D environment. Besides that, previous approaches that similar and relevant to this study for 3D environment only focus on specific type of 3D environment [7]. We also aim to make our method flexible that can be implemented in any real world environment.

The purpose of our research is to find the minimum number of the cameras and the positions for the camera within the 3D environment such that the *coverage* of the camera is maximised. Our approach focuses on observing the interior of the 3D environment with or without objects. Given a 3D environment and a set of cameras, we want to find the placement for the cameras as well as the minimum number of the cameras needed for that environment. We intend to make the camera placement system relevant for any purpose therefore we try to avoid any constraint other than the dimension of the 3D environment, the obstacles inside it and the FOV of the camera. Various sizes and positions of the 3D objects (obstacles) are used as the sample for testing to find the most accurate approach for the camera placement system. Industry environment such as oil plant or manufacturing factory also is used as the testing environment. This is to make sure the system developed is applicable to the real world environment.

The next section for this paper is organised as follows: we discuss related work in Section 2 and in Section 3 we present our methodology to solve the camera placement problem. Section 4 shows our experimental result and Section 5 has our conclusions and future plan.

2 Related Works

Most of previous camera placement approaches were proposed based on original Art Gallery Problem and implemented in two-dimensional (2D) environment. For 2D environment, grid mapping approach has been applied as in [8],[9],[10] and [11] where the camera locations are calculated on the coverage map. This coverage map is represented by a set of grids in order to convert the problem into the discrete domain [8]. In [8], ‘feasible region’ on the grid map is considered as the intersecting area of visibility constraint, resolution constraint and camera FOV. E. Horster *et al.* [10] proposed sampling a space according to various constraints as an alternative of using a regular grid. Another method for 2D environment was proposed in [12] which is landmark-based approach. The authors modified common sensor coverage approach in which case, they changed the concept of sampling a set candidate sensor placements to a set of landmarks that should be observed. In [11], they only concerned about space of target locations, camera directions, camera locations and the target perimeter.

Current work focuses on 3D environment and the previous placement system proposed in 2D can be enhanced into 3D as shown in [13] in order to overcome the limitations in 2D. Optimal sensor position in [13] is determined based on the intersecting line of active vertexes and surfaces (regions) of the objects inside specific 3D environment. Previous work only tested using some simple cases and improvement for their work can be done for all type of general cases or environment. E. Becker *et al.* [7] and A. Van *et al.* [14] also implement their camera placement methods in 3D. In [7], given a 3D environment, volume of interest consists of a set of points is calculated. These points are responsible to vote in the voting scheme method to find out which sensors are observing the points at the current time.

Basically flow of proposed camera placement algorithm that has been applied in previous works is similar to [15] and [16]. As shown in previous works, visibility test is important in order to see the actual coverage the camera FOV. Both camera specification in [15] and [16] has finite FOV but in [16], the authors did not consider depth of field of the sensor as one of the constraints. Based on the research done, there are three types of camera used in surveillance and monitoring system which are directional camera, omnidirectional camera and PTZ camera. Previous works that have addressed coverage problems for directional sensors are [10], [17] and [18]. J. Gonzalez *et al.* [20] used the combination of directional and omnidirectional cameras to solve the optimization problem. They also used grid map, same like in [8]. We exclude the type of camera from our parameter as we only use fixed camera or directional camera for this research. Triangulation based is also one of the methods for camera placement system [21]. This method required measurements from the two or more connected sensors to be combined for an estimation of the position. Using the combination of two sensors the uncertainty in estimating the position of the target placement can be minimised. They also address the occlusion constraint which is not calculated in the 'Art Gallery Problem'.

E. A. Sultanik *et al.*[22] introduced a distributed version of a multiagent approximation algorithm for solving the distributed art gallery and dominating set problem. The dominating set problem involves the connectivity problem to find a minimum subset of the vertices on a polygon such that every vertex not in the subset has at least one member of the subset in its neighbourhood [22]. Another previous work that proposed distributed approach or decentralised approach is discussed in [23]. In previous paper, the number of the guards is fixed and the number of guarded painting is maximised instead of finding the minimum number of guards.

In [4] the authors stated the theorem that for a polygon with holes, the approximate solution of minimum vertex guard can be computed using $O(n^5)$ and the theorem for egde guard is $O(n^4)$. D. Jang *et al.* [3] proposed a restricted version of original art gallery problem known as point guard problem [3] and their approach is improved from previous algorithm in [4]. The vertex guard and the edge guard must be visible from some edges and vertices in each respective set. The authors present two approximation algorithms each one for vertex guards problem and edge guards problem [3]. Both algorithms show that the optimal solution can be calculated with $O(n^3)$ for any simple polygon.

Most of previous papers focus on 2D placement that only deal with width and depth of particular space. With limited parameters the results produced from 2D placement method are not accurate and applicable to real world. Hence, our study focuses on 3D placement that takes the height of the space as one of the main parameters to compute the space's volume. In this study we use volume approach that makes use of camera FOV volume and space's volume to calculate the minimum number of cameras needed to observe particular space.

3 Methodology

Our research objective is to find the minimum number of cameras and their placement in 3D environment. Specific requirements are gathered in order to achieve the objective of this study. These requirements are the coverage of the camera, visibility, limitations of the camera specification and the size and dimension of the environment layout. The camera specifications used in this research has limited vertical and horizontal viewing angle to follow the specification of some of the cameras used in real world. In our approach we use a frustum to represent the camera FOV. Frustum shape is selected because it has different angle for horizontal side and vertical side and similar to real camera FOV as in Fig. 1.

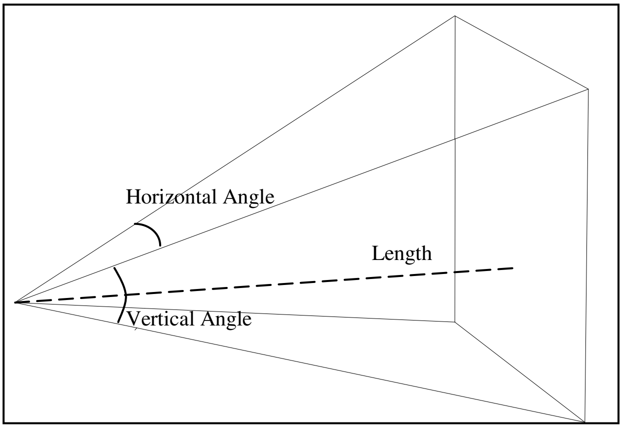


Fig. 1. Frustum as the camera FOV

For this research, we will implement the proposed algorithm inside existing software to test it. As stated earlier, in our approach we try to avoid any constraints other than the dimension of the environment and 3D objects inside the environment (occluders) to make sure the placement system is flexible and reliable. It will be more useful if the placement system can be used for any type of environment rather than to focus on one type of environment. We also consider that our approach to be promisingly applicable to the real world environment. The flow of our approach is shown in Fig. 2.

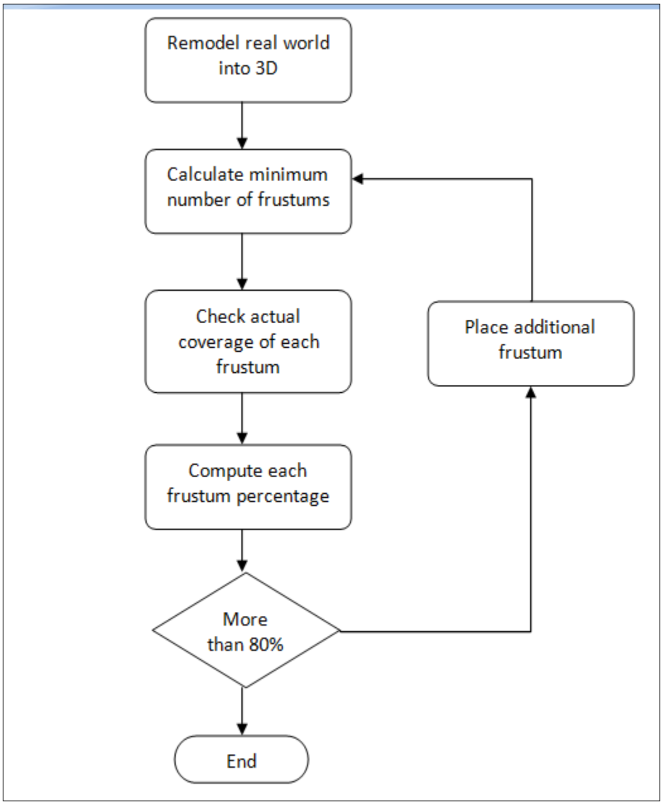

Remodel real world into 3D
Calculate minimum number of frustums
Check actual coverage of each frustum
Compute each frustum percentage
More than 80%
End
Place additional frustum

Fig. 2. Flowchart of proposed system

The first step is to remodel the selected real world environment into 3D environment. To ensure this system is applicable in real world application, all the measurement and the coordinates of the 3D models is based on the exact values (meters). Based on the calculated volume of the 3D environment, number of minimum frustum is computed as we already know the exact volume for one frustum. Eq. 1 gives the calculated volume

$$fVolume = \frac{\text{space's volume}}{\text{frustum's volume}} \quad . \tag{1}$$

Then, frustums will be placed inside the environment to indicate the cameras. Because we plan to make the system automated, we take the input of minimum number of frustums and place the frustums automatically inside the environment. All the frustums must be placed on the wall and facing towards the interior of the environment. In some cases, not all places inside the 3D space need to be observed and the important part that must be observed is considered as the target region. Despite having the target region, the placement system of the camera still focuses to maximise the space of the coverage. Transparent 3D cube is placed to indicate the target region and its size depends on the users input. Based on its size, number of the cameras that monitor the target is calculated.

Frustum works like a light source which means its volume will change if part of it is occluded by any of the 3D objects inside the environment. An object inside the 3D environment is said to be visible from the source of the frustum if any line joining point on the object and the source does not intersects other objects and lies inside the frustum. The occluded part can be considered same as the shadowed region which is the region that cannot be seen from a camera position. To ensure particular space of the 3D environment is covered by at least one frustum, the occluded part need to be eliminated. This process is done to show the actual coverage of the frustum. The nearest obstacle inside the frustum will be calculated first for the occluded part and then that particular part is eliminated before the same process repeats for the next obstacles. The process only repeats if the next obstacles are still inside the frustum. This step is done to minimise the computational load and time as the part that already being eliminated is not included in the next calculation. This will guarantee that the system would not be complex and involve difficult computation.

After the previous process completed, the percentage of the frustums is calculated to see whether the coverage is sufficient for that particular environment or not. We set 80% as the limit for checking the percentage. If the coverage of each frustum is less than 80% the system will place additional frustum at particular space. The percentage of the coverage is calculated from the volume of the actual frustum coverage after the occluded part has been eliminated. As the shape of the final frustum will become irregular we use Surveyor's formula (Eq. 2) to determine the internal area of the geometry [24]. From the internal area, the volume of each frustum can be computed.

Surveyor's formula

$$area = \frac{1}{2}\sum\nolimits_{k=1}^{n} \begin{vmatrix} \mathrm{x}_{\mathrm{k}-1} & \mathrm{x}_{\mathrm{i}} \\ \mathrm{y}_{\mathrm{k}-1} & \mathrm{y}_{\mathrm{i}} \end{vmatrix} . \tag{2}$$

4 Result and Discussion

To check our method to form the actual coverage of the frustum in 3D environment we use simple test case with one frustum (45° horizontal angle and 45° vertical angle) and one object. A frustum is placed inside 3D environment with a cuboid as shown in Fig. 4 and Fig. 5. 2D view of the test case is shown in Fig. 3 which only shows the top view of the objects and the frustum is assumed to cover half of the objects. However, in Fig. 4 we can clearly see only quarter of the object is covered by the frustum. This is one of the limitations in 2D approach where the actual space that been observed is not accurate.

To calculate the actual coverage of the frustum, our system will check each one of the frustum inside the environment and the objects within the frustum coverage. In 2D perspective, the occluded part between the frustum and the cuboid can be identified as the shadow part. Similar in 3D environment, the occluded part is formed from the projection of each vertex on the cuboid onto the plane (front surface) of the frustum. The source point to compute the shadow part is the source point of the frustum. 3D shadow formed then will be subtracted from the original frustum to produce actual coverage as shown in Fig. 6. For this study, the actual coverage of a frustum is computed in order to maximise coverage of the camera.

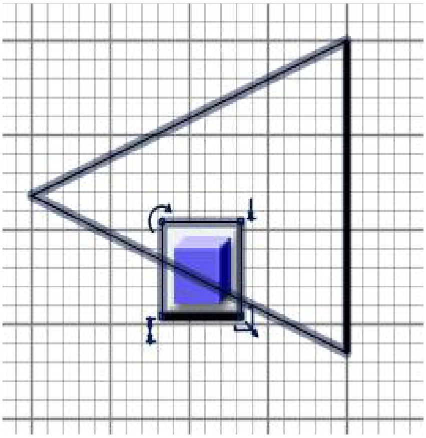


Fig. 3. 2D Workspace top view with one object within the frustum coverage

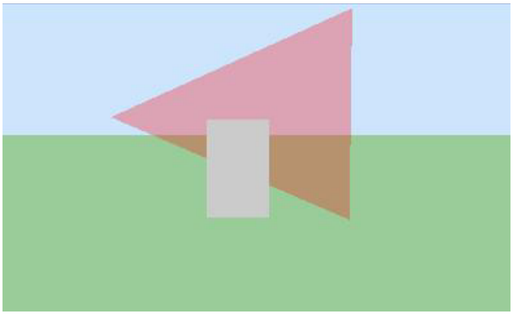


Fig. 4. Side view of the previous image in 3D environment

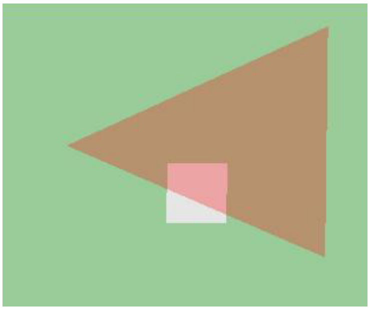


Fig. 5. Top view in 3D environment

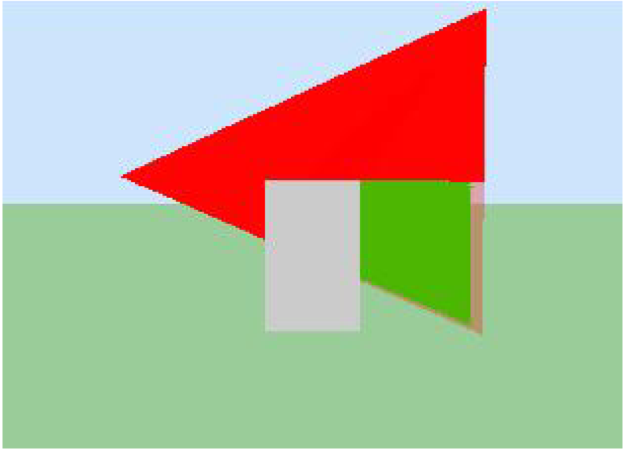


Fig. 6. Side view of actual frustum coverage

For another test case without holes (objects) we present experimental results about the comparison between previous 2D camera placement [25] and the actual coverage in 3D environment. We use same polygon layout in Fig. 7. and remodelled the layout into 3D workspace as shown in Fig. 8. Purposely for this test case, the viewing angle used for testing are fixed at two values which are 90° for horizontal angle and 60° for vertical angle. The length of the frustum will be the distance from the wall to another wall or from wall to the target region. We make the position of a camera to be only on wall as the camera used is a static directional camera. We chose staircase polygon case from [25] because they assumed the camera FOV to be 90° and the camera position must be on the vertices of the polygon which is in our case to be on the wall not in the interior of the polygon.

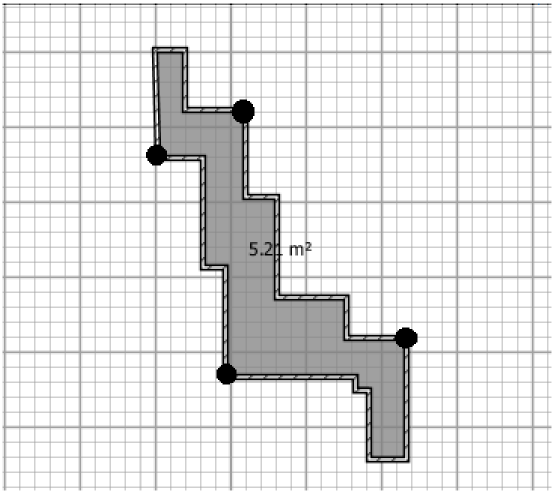


Fig. 7. Polygon layout and guard placement taken from [25]

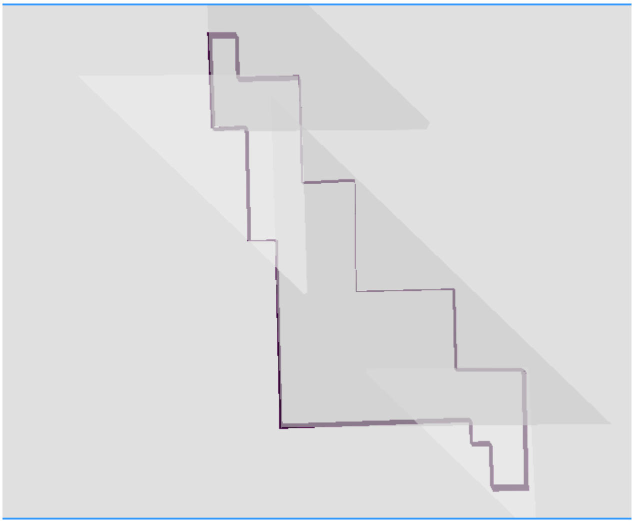


Fig. 8. Remodelled previous figure into 3D

Table 1 shows the coverage for 2D case while Table 2 shows the coverage for 3D case. Coverage for each case is compared in term of volume (m^3) to see the differences between 2D and 3D placement. As in Table 2, the average percentage of the coverage is 66.65% which means the coverage is insufficient for that space.

Table 1. Result For 2D Placement Coverage

	Coverage(m^3)
Camera 1	304.3116
Camera 2	702.144
Camera 3	296.6558
Camera 4	139.5633
Total Coverage	1442.675

Table 2. Result for 3D Placement Coverage

	Coverage (m^3)	Coverage %
Camera 1	202.79	66.60%
Camera 2	468.12	66.67%
Camera 3	197.7753	66.67%
Camera 4	93.0417	66.67%
Total Coverage	961.727	

The actual coverage of the camera placement in 3D space is shown in Fig.9 and the blind spot of the camera or the place below the camera FOV is not covered at all. As in 2D we cannot see the actual height of the camera FOV because previous work assumed the vertical angle of the camera to be 180^o. To solve this problem we placed another camera facing towards the other camera to cover its blind spot as shown in Fig. 10. From this result it is clear that the location of the camera itself should be covered by at least one camera to make sure maximum coverage.

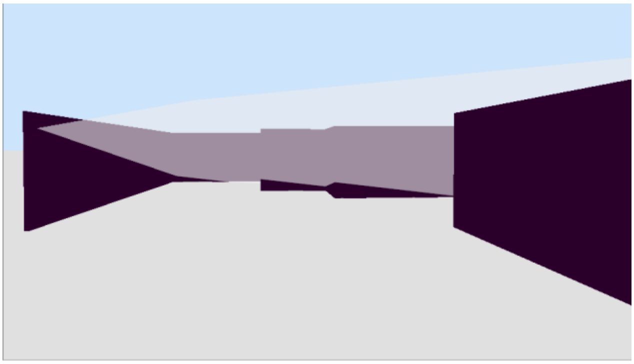


Fig. 9. Actual coverage seen from the side view

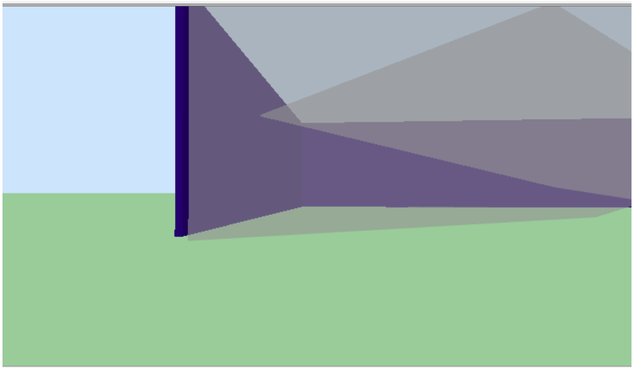


Fig. 10. Another frustum covers the original frustum

The main important advantage of our system is that the actual coverage of the cameras, the vertical angle of camera FOV and the objects inside it can be measured and seen. This is also the main purpose why we implement the camera placement in 3D. Most of the previous works only implemented in 2D, where they assumed the camera has 180^o vertical angle. Using our 3D system, vertical angle of the camera can be seen and calculated to get the actual coverage of the space. This will optimise the placement of the cameras because we can manipulate the whole volume of 3D environment.

5 Conclusions

We have developed a brief flow of methodology for camera placement system for 3D environment and our method is based on the interaction with 3D models. The intersected part between the coverage of the frustum and occluded part inside that space is subtracted from the frustum. From this we are able to see the actual coverage of the frustum and maximise the observed space. For future work, we plan to enhance our program so that it can handle the automation for camera placement and more complex 3D environment and several constraints.

References

1. Urrutia, J.: Art Gallery and Illumination Problems, 973–1027 (2000)
2. Couto, M. C., de Rezende, P. J., de Souza, C. C.: An IP Solution to the Art Gallery Problem, 88–89 (2009)
3. Jang, D., Kwon, S.: Fast Approximation Algorithms for Art Gallery Problems in Simple Polygons. CoRR, abs/1101.1346 (2011)
4. Ghosh, S.K.: Approximation Algorithms for Art Gallery Problems in Polygons. Discrete Applied Mathematics 158, 718 (2010)
5. Fragoudakis, C., Markou, E., Zachos, S.: Maximizing the Guarded Boundary of an Art Gallery is APX-Complete. Comput.Geom.Theory Appl. 38, 170–180 (2007)
6. Tóth, C.D.: Art Galleries with Guards of Uniform Range of Vision. Computational Geometry 21, 185–192 (2002)
7. Becker, E., Guerra-Filho, G., Makedon, F.: Automatic Sensor Placement in a 3D Volume. 36, 1–36 (2009)
8. Erdem, U.M., Sclaroff, S.: Automated Camera Layout to Satisfy Task-Specific and Floor Plan-Specific Coverage Requirements. Comput.Vis.Image Underst. 103, 156–169 (2006)
9. Zou, Y., Chakrabarty, K.: Sensor Deployment and Target Localization in Distributed Sensor Networks. ACM Trans. Embed. Comput. Syst. 3, 61–91 (2004)
10. Horster, E., & Lienhart, R.: On the Optimal Placement of Multiple Visual Sensors, pp. 111–120 (2006)
11. Mostafavi, S.A., Dehghan, M.: Optimal Visual Sensor Placement for Coverage Based on Target Location Profile. Ad Hoc Network (in Press, Corrected Proof)
12. Agarwal, P.K., Ezra, E., Ganjugunte, S.K.: Efficient Sensor Placement for Surveillance Problems. In: Krishnamachari, B., Suri, S., Heinzelman, W., Mitra, U. (eds.) DCOSS 2009. LNCS, vol. 5516, pp. 301–314. Springer, Heidelberg (2009)
13. Bottino, A., Laurentini, A.: Optimal Positioning of Sensors in 3D. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 804–812. Springer, Heidelberg (2005)
14. Van den Hengel, A., Hill, R., Ward, B., et al.: Automatic Camera Placement for Large Scale Surveillance Networks. In: Workshop on Applications of Computer Vision (WACV), pp. 1–6 (2009)
15. Yabuta, K., Kitazawa, H.: Optimum Camera Placement Considering Camera Specification for Security Monitoring, pp. 2114–2117 (2008)
16. Debaque, B., Jedidi, R., Prévost, D.: Optimal video camera network deployment to support security monitoring. In: 12th International Conference on Information Fusion, FUSION 2009, pp. 1730–1736 (2009)
17. Fusco, G., Gupta, H.: Placement and Orientation of Rotating Directional Sensors. In: SECON, pp. 332–340 (2010)

18. Fusco, G., Gupta, H.: Selection and Orientation of Directional Sensors for Coverage Maximization. In: 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON 2009, pp. 1–9 (2009)
19. David, P., Idasiak, V., Kratz, F.: A Sensor Placement Approach for the Monitoring of Indoor Scenes. In: Kortuem, G., Finney, J., Lea, R., Sundramoorthy, V., et al. (eds.) EuroSSC 2007. LNCS, vol. 4793, pp. 110–125. Springer, Heidelberg (2007)
20. Gonzalez-Barbosa, J., Garcia-Ramirez, T., Salas, J., et al.: Optimal Camera Placement for Total Coverage, pp. 844—848 (2009)
21. Tekdas, O., Isler, V.: Sensor Placement Algorithms for Triangulation Based Localization. In: IEEE International Conference on Robotics and Automation, 364164, pp. 4448–4453 (2007), doi:10. 1109/ROBOT
22. Sultanik, E. A., Shokoufandeh, A., Regli, W. C.: Dominating Sets of Agents in Visibility Graphs: Distributed Algorithms for Art Gallery Problems, pp. 797–804 (2010)
23. Lass, R.N., Grauer, M.J., Sultanik, E.A., Regli, W.C.: A Decentralized Approach to the Art Gallery Problem (2007)
24. B.B.: The College Mathematics Journal 17, 326–337 (1986)
25. Erickson, L.H., LaValle, S.M.: A chromatic art gallery problem. University of Illinois at Urbana-Champaign, Urbana, IL, Tech. Rep. (2010)

XRecursive: AStorage Method for XML Document Based on Relational Database

M.A. Ibrahim Fakharaldien, Jasni Mohamed Zain, and Norrozila Sulaiman

Faculty of Computer System and Software Engineering,
University Malaysia Pahang,
Kuantan, Malaysia.
mohfakhrdeen@gmail.com, jasni@ump.edu, norrozila@ump.edu

Abstract. Storing XML documents in a relational database is a promising solution because relational databases are mature and scale very well and they have the advantages that in a relational database XML data and structured data can coexist making it possible to build application that involve both kinds of data with little extra effort. In this paper, we propose an algorithm schema named XRecursive that translates XML documents to relational database according to the proposed storing structure. The steps and algorithm are given in details to describe how to use the storing structure to storage and query XML documents in relational database. Then we report our experimental results on a real database to show the performance of our method in some features.

Keywords: XML, Relational Database, SQL.

1 Introduction

Today's data exchange between organizations has become challenging because of the difference in data format and semantics of the meta-data which used to describe the data. Now day' XML emerged as a major standard for representing data on the World Wide Web while the dominant storage mechanism for structured data is the relational databases, which has been an efficient tool for storing, searching, retrieving data from different collection of data. The ability to map XML data in relational databases is difficult mission and challenging in the world of all IT organization so there is a need to develop an interfaces and tools for mapping and storing XML data in relational databases.

The extensible Markup Language (XML) is quickly becoming the de facto standard for data exchange over the Internet [10] and now it plays a central role in data management, transformation, and exchange. Since its introduction to industry in the Late 1990s, XML [1] has achieved widespread support and adoption among all the leading software tools, server, and database vendor s. As importantly, XML has become the lingua franca for data by lowering the cost of processing, searching, exchanging, and re-using information. XML provides a standardized, self-describing means for expressing information in a way that is readable by humans and easily verified, transformed, and published, the hot topic is to seek the best way for storing XML documents in order to get high query processing efficiency[12]. In addition,

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 460–469, 2011.

data can be transmitted to remote services anywhere on the Internet using XML-based Web services to take advantage of the new ubiquity of connected software applications. The openness of XML [2] allows it to be exchanged between virtually any hardware, software, or operating system. Simply put, XML opens the door for information interchange without restriction. Today, the dominant storage mechanism for structured enterprise data is the relational database, which has proven itself an efficient tool for storing, searching for, and retrieving information from massive collections of data. Relational databases specialize in relating individual data records grouped by type in tables. Developers can join records together as needed using SQL (Structured Query Language) and present one or more records to end-users as meaningful information. The relational database model revolutionized enterprise data storage with its simplicity, efficiency, and cost effectiveness. Relational databases have been prevalent in large corporations since the 1980s, and they will likely remain the dominant storage mechanism for enterprise data in the foreseeable future. Despite these strengths, relational databases lack the flexibility to seamlessly integrate with other systems, since this was not historically a requirement of the database model [3]. In addition, although relational databases share many similarities, there are enough differences between the major commercial implementations to make developing applications to integrate multiple products difficult. Among the challenges are differences in data types, varying levels of conformance to the SQL standard, proprietary extensions to SQL, and so on.For the storage of XML document, the key issue is transforming the tree structure of an XML document into tuples in relational tables [11].Nowadays, there are more and more data presented as XML document, the need of storing them persistently in a database has increased rapidly while the native–XML databases usually have limited support for relational databases. In recent years, with the popularity of relational databases (RDB), approaches based on RDB[4,5,6,7,8,9,] to store and manipulate XML data as relational tables but still there is need to manage XML data and relational data seamlessly with similar storage and retrieval efficiencies simultaneously. XML and Relational databases cannot be kept separately because XML is becoming the universal standard data format for the representation and exchanging the information whereas most existing data lies in RDBMS and their power of data capabilities cannot be degraded so the solution to this problem a new efficient methods for storing XML documents in relational database is required.A new efficient method for storing XML document in relational database is proposed in this paper to face these problems.

The rest of the paper is organized as follows. Section 2 briefly discusses existing techniques to store and query XML in an RDBMS. The database schema of XRecursiveis presented in Section 3 which briefly describe how an XML document is stored in an RDBMS using XRecursive. InSection 4, we present the analysis of the experimental results. The last section concludes the paper.

2 Related Works

There are basically three alternatives for storing XML data: in semi-structured databases [15], in object-oriented databases [13], and in relational systems [21–19, 17, 14, 22, 16, and 23]. Among these approaches, the relational storage approach has attracted considerable interest with a view to leveraging their powerful and reliable data management services. In order to store an XML document in a relational database, the tree structure of the XML document must first be mapped into an equivalent, flat, and relational schema. XML documents are then shredded and loaded into the mapped tables. Finally, at runtime, XML queries are translated into SQL, submitted to the RDBMS, and the results are then translated into XML. There is a rich literature addressing the issue of managing XML documents in relational back- ends [21–19, 17, 14, 22, 16, and 23]. These approaches can be classified into two major categories as follows:

1.1 Schema-conscious approach: This method first creates a relational schema based on the DTD/schema of the XML documents. First, the cardinality of the relationships between the nodes of the XML document is established. Based on this information a relational schema is created. The structural information of XML data is modeled by using primary-key foreign-key joins in relational databases to model the parent–child relationships in the XML tree. Examples of such approaches are Shared-Inlining [14], LegoDB [21, 20]. Note that this approach depends on the existence of a schema describing the XML data. Furthermore, due to the heterogeneity of XML data, in this approach a simple XML schema/DTD often produce a relational schema with many tables.

1.2 Schema-oblivious approach: This method maintains a fixed schema which is used to store XML documents. The basic idea is to capture the tree structure of an XML document. This approach does not require existence of an XML schema/DTD. Also, number of tables is fixed in the relational schema and does not depend on the structural heterogeneity of XML documents. Some examples of schema-oblivious approaches are Edge approach [19], XRel [16], XParent [17].Schema-oblivious approaches have obvious advantages such as the ability to handle XML schema changes better as there is no need to change the relational schema and a uniform query translation approach. Schema-conscious approaches, on the other hand, have the advantage of more efficient query processing [18]. Also, no special relational schema needs to be designed for schema-conscious approaches as it can be generated on the fly based on the DTD of the XML document(s).

3 The Proposed Method

3.1 XML Document

The data structure of XML document is hierarchical, consist of nested structures. The elements are strictly marked by the beginning and ending tags, for empty elements by empty-elements tags. Character data between tags are the content of the elements. It is an instance of XML document contains information about an employee as follows.

```
<? Xml version="1.0" encoding="UTF-8"?>
<Personnel>
<Employee type="permanent">
<Name>Seagull</Name>
<Id>3674</Id>
<Age>34</Age>
</Employee>
<Employee type="contract">
<Name>Robin</Name>
<Id>3675</Id>
<Age>25</Age>
</Employee>
<Employee type="permanent">
<Name>Crow</Name>
<Id>3676</Id>
<Age>28</Age>
</Employee>
</Personnel>
```

Fig. 1. XML Document

3.2 The Tree Structure Representation of XML Document

In this section, the structure independent mapping approach is explained with a sample XML document shown in above Figure 1.

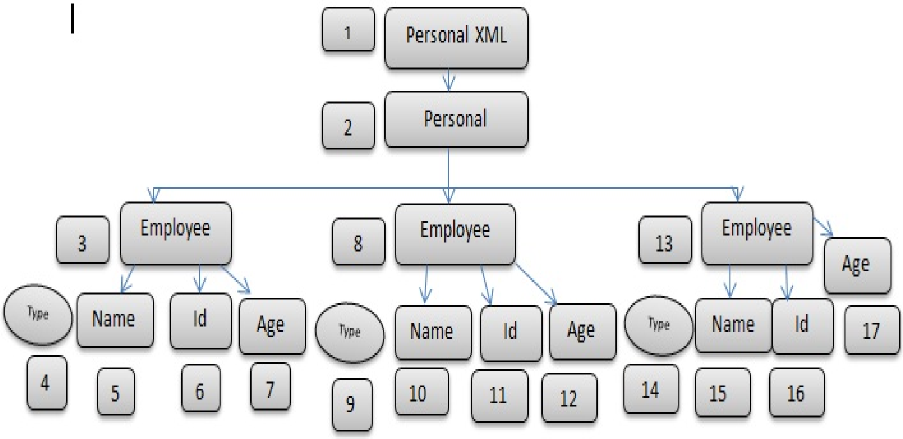


Fig. 2. Tree structure of XML document with XRecursive labeling

3.3 XRecursive Structure

Each and every XML can be describing as a XML tree. In this figure the squares are the elements and the ovals are the attributes of the elements. A generated XML tree has been shown in the figure. Every element or attributes are identified by a signature (number).

Definition 1. *XRecursive Structure: XRecursive is a storage structure for storing XML documents whereeach path is identified by its parent from the root node in a recursive manner.*

3.4 Algorithm

XML document can be stored in relational database, in this paper, MYSQL by use of above two tables. In this paper algorithms are proposed to store XML document into relational database as the following:

```
Algorithm: store_to_DB(xml as a document)

1: Begin
2: let N = { } as empty set, where N represents the list of node of the XML file.
3. Let V = { } as empty set, where V represents the list of the value of the node.
4. Let String filename = null and int id = 1;
5: filename = read XML document name
6: while xml Document is not null
7: Begin
8:     read element name as name
9:     read the element type
10:    if element type is Element or attribute
11:    begin
12:        read parent as pName
13:        id = id + 1;
14:        add name, pName, id to the N
15:    End
16:    else if element type is #text
17:    Begin
18:        Read textvalue as value
19:        Id = id + 1;
20:        Add name, value, id to V
21:    end
22: store N into database
23: store V into database
24: End.
```

Example 1. In this structure when an element or type associates with its signature it also represents its parent element. We add document name in association with the id to be able to add multiple XML file in the storage. Figure 2 represents the storage of the XML file associated with its signature. For every element there will have a signature associated with it and there will also have a parent's signature associated with it. In table 1: tagName represents the name of the node; id represents the id of the node which is the PK. And finally pId represents the parent id of the node. As document name don't have any parent id so the id of the document name and parent id of the document name is same that has been shown in the figure 2.

Table 1. Tag_structure

tagName	Id	pId
Personal.xml	1	1
personal	2	1
Employee	3	2
type	4	3
name	5	3
id	6	3
age	7	3
Employee	8	2
type	9	8
name	10	8
id	11	8
age	12	8
Employee	13	2
type	14	13
name	15	13
id	16	13
age	17	13

Table 2. Tag_value

tagId	Value	Type
4	Permanent	A
5	Seagua11	E
6	3674	E
7	34	E
9	Contract	A
10	Robin	E
11	3675	E
12	25	E
14	Permanent	A
15	Crow	E
16	3676	E
17	28	E

In table 2, we represent the value associated with the elements or type. In XRecursive structure there is no need to store the path value or path structure as it will be determine recursively by its parent id. In Table 1: tagName is the name of the tag, where Id is the parent key. In Table 2: tagId presents the Table 1 id thus tagIdis the foreign key. In Table 2 tagId only represents the elements which contain a value and the value represents on the value column. And the type ‘A’ denoted to the attribute and ‘E’ denoted to the element.

4 The Analysis of Experiment

We ran experiments using our xml document in Fig.1. Our experiment was performed on 3.00 GHz Pentium 4 processor with 4GB RAM, 240 GB of hard disk running on windows XP system, we used MySQL v5.01 as the database for storing XML document and java language (Version jdk6) for parsing the XML document and then save it in MySQL . The experiments are conducted using the XML benchmark. The experiment evaluates the efficiency of storing XML document in relational database based-on the XRecursive structure. The experiment is made with respect to the following four factors:

a. Database Size:
XRecursive has lesser storage requirement than the SUCXENT, as XRecursive only uses two tables to store the XML data whereas SUCXENT uses five tables. Comparison can be seen from the fig.3 which shows that by this storage method we can reduce not only the size of database requirement of the labeling of node, but also the number of tables.

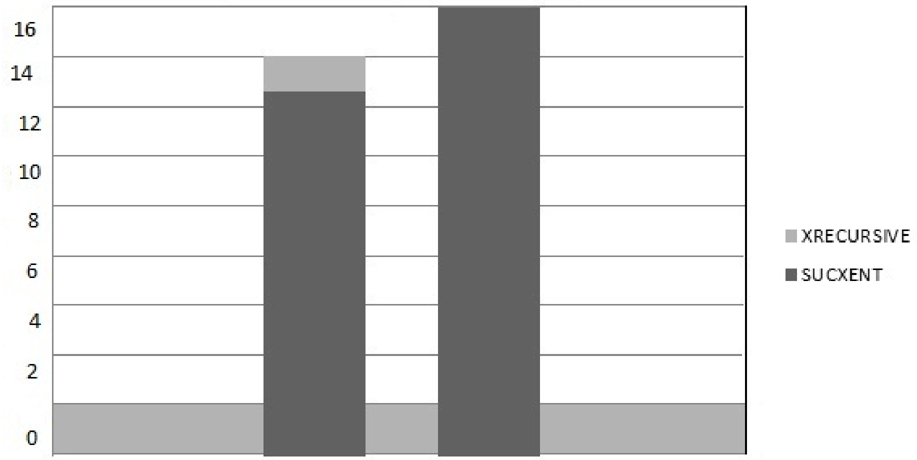


Fig. 3. Database Size in MB

b. Insertion Time
The Fig. 4 Shows comparison of the document's insertion time. XRecursive method is approximately 1.5 times faster than the SUCXENT method. The reason because XRecursive only uses two tables whereas SUCXENT uses five tables which requires some more processing time during the filing the contents of tuples in each table.

c. Retrieval Time
The document retrieval time is given in Fig 5. The results are much closed. XRecurisve seems to be little faster than SUCXENT.

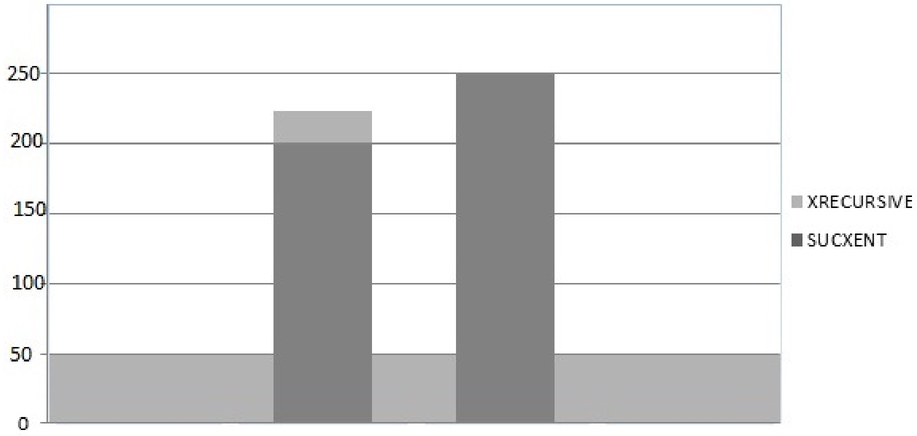


Fig. 4. Insertion Time in Second

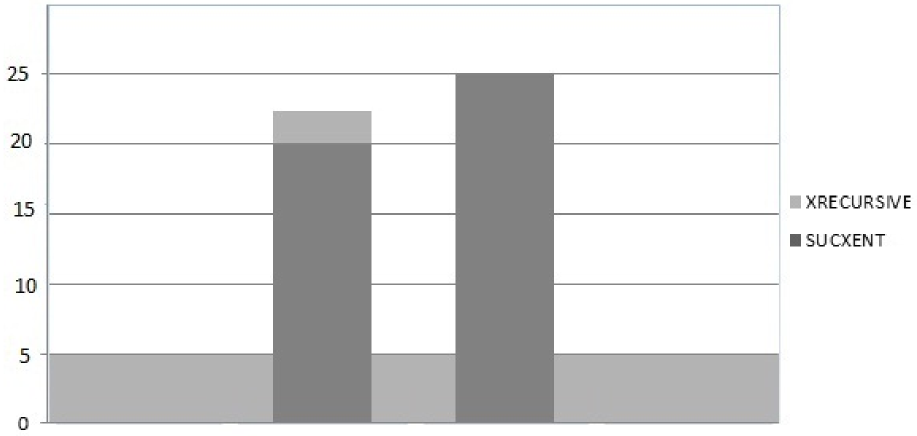


Fig. 5. Retrieval Time in Second

5 Conclusion

XRecursive, a general storage method for XML document using relational database is proposed in this paper. XRecursive adopts the model-mapping method to store XML document in relational database, to decompose the tree structure into nodes and store all information of nodes in relational database according to the node types by recursive way. It can deal with any documents no matter whether it has fixed schema or not. By using this method we can reduce the database size require to store the XML document into relational database. The storing algorithm of XML document into relational database was also given in the paper, and examined the accuracy of it by using the XML document in performance section. Utilizing the actual Xml document evaluated the performance of storing XML document into relational database by using our method.

References

1. Grandi, F., Mandreoli, F., Tiberio, P., Bergonzini, M.: A temporal data model and management system for normative texts in XML format. In: Proceedings of the 5th ACM international Workshop on Web information and Data Management, New Orleans, Louisiana, USA, November 07 - 08 (2003)
2. Augeri, C.J., Bulutoglu, D.A., Mullins, B.E., Baldwin, R.O., Baird, L.C.: An analysis of XML compression efficiency. In: Proceedings of the Workshop on Experimental Computer Science, San Diego, California (June 2007)
3. Reed, D.: Take a good look. Data Strategy, from Business Source Complete database 2(4), 24–29 (2008)
4. Sybase Corporation: Using xml with the Sybase adaptive server sol databases. Technical whitepaper(August 21, 1999)
5. XRel: a path-based approach to storage and retrieval of xml documents using relational databases. ACM Trans. Interet. Technol. 1(1), 110–141 (2001)
6. Xparent: An efficient rdbms-based xml database system. In: ICDE 2002: Proceedings of the 18th International Conference on Data Engineering, p. 335. IEEE Computer Society, Washington, DC, USA (2002)
7. Jiang, H., Lu, H., Wang, W., Yu, J.X.: Path materialization revisited: an efficient storage model for xml data. In: ADC 2002: Proceedings of the 13th Australasian database conference, pp. 85–94. Australian Computer Society, Inc. (2002)
8. Kyung-Soo, J.: A design of middleware components for the connection between xml and rdb. In: Proceeding of the IEEE International Symposium on Industrial Electronics, pp. 1753–1756 (2001)
9. Rys, M.: Microsoft sol server 2000 xml enhancements. Microsoft Support Webcast (April 2000)
10. HasanZafari, K.M.: EbrahimShiri. Xlight, An Efficient Relational Schema To Store And Query XML Data. In: Proceeding of the IEEE Internationalconference inData Store and Data Engineering, April 22, pp. 254–257 (2010)
11. Yue, L., Ren, J., Qian, Y.: Storage Method of XML Documents Based-on Pri-order Labling Schema. In: Proceeding of the IEEE International Workshop on Education Technology and Computer Science, December 30, pp. 50–53 (2008)
12. Sainan, L., Caifeng, L., Liming, G.: Storage Method for XML Document based on Relational Database. In: Proceeding of the IEEE International Symposium on Computer Science and Computational Technology, pp. 127–131 (2009)
13. Bancihon, F., Barbedette, G., Benzaken, V., et al.: The design and implementation of an object-oriented database system. In: Proceedings of the Second International Workshop on Object-oriented Database (1988)
14. Shanmugasundaram, J., Tufte, K., et al.: Relational databases for querying XML documents: limitations and opportunities. In: VLDB (1999)
15. Goldman, R., McHugh, J., Widom, J.: Fromsemi structured data to XML: migrating the lore data model and query language. In: Proceedings of WebDB 1999, pp. 25–30 (1999)

16. Yoshikawa, M., Amagasa, T., Shimura, T., Uemura, S.: XRel: a path-based approach to storage and retrieval of xmldocuments using relational databases. ACM TOIT 1(1), 110–141 (2001)
17. Jiang, H., Lu, H., Wang, W., Xu Yu, J.: Path materialization revisited: an efficient storage model for XML data. In: 13th Australasian Database Conference, ADC (2002)
18. Tian, F., DeWitt, D., Chen, J., Zhang, C.: The design and performance evaluation of alternative XML storage strategies. ACM Sigmod Record (2002)
19. Florescu, D., Kossman, D.: Storing and querying XML data using an RDBMS. IEEE Data Engineering Bulletin (1999)
20. Ramanath, M., Freire, J., Haritsa, J., Roy, P.: Searching for efficient XML-to-relational mappings. In: Proceedings of the International XML Database Symposium (2003)
21. Bohannon, P., Freire, J., Roy, P., Simeon, J.: From XML schema to relations: a cost-based approach to XMLstorage. In: Proceedings of IEEE ICDE (2002)
22. Tatarinov, I., Viglas, S., Beyer, K., et al.: Storing and querying ordered XML using a relational database system. In: Proceedings of the ACM SIGMOD (2002)
23. Zhang, C., Naughton, J., Dewitt, D., Luo, Q., Lohmann, G.: On supporting containment queries in relational database systems. In: Proceedings of ACM SIGMOD (2001)

Enhancement Algorithms for SQL-Based Chatbot

Abbas Saliimi Lokman* and Jasni Mohamad Zain

Faculty of Computer Systems & Software Engineering,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Kuantan, Pahang Darul Makmur
abbassaliimi@yahoo.com,
jasni@ump.edu.my
http://www.ump.edu.my

Abstract. Artificial intelligence chatbot is a technology that makes interaction between men and machines using natural language possible. From literature of chatbot's keywords/pattern matching techniques, some potential issues for enhancement had been discovered. The discovered issues are in the context of relation between previous and next responses/outputs and also keywords arrangement for matching precedence together with keywords variety for matching flexibility. To encounter these issues, two respective algorithms had been proposed. Those algorithms are *Extension and Prerequisite* and *OMAMC (One-match and All-match Categories)*. Implemented in SQL-Based chatbot, both algorithms are shown to be enhancing the capability of chatbot's keywords/pattern matching process by providing an augment ways in storing the data and performing the process. This paper will present the significance of results from implementing both proposed algorithms into SQL-Based chatbot that will result on some enhancements in certain area of chatbot's processes.

Keywords: chatbot, keywords/pattern matching, response relation, keywords category, matching precedence.

1 Introduction

In 1950, mathematician Alan Turing proposed the question "Can machines think?" [13]. Since then, a number of attempts to tackle this question had been emerged in computer science field that later formed the field of Artificial Intelligence. One of many attempts to visualize an intelligence machine is chatbot or chatter robot. Chatbot is a technology that enabled an interaction between man and machine using human natural language. First introduced by Weizenbaum (an MIT professor) in 1966 [15], the chatbot named ELIZA later became a main inspiration for computer science and linguistic researchers in creating an

[*] We would like to thank Universiti Malaysia Pahang (UMP) for the support grant for this research no. RDU100360.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 470–479, 2011.

application that can understand and response to human natural language. The huge breakthrough in chatbot technology came in 1995 when Dr. Richard Wallace, an ex-Professor of Carnegie Mellon University combined his background in computer science with his interest in internet and natural language processing to create Artificial Linguistic Internet Computer Entity or A.L.I.C.E. [14]. A.L.I.C.E. that later being described as modern ELIZA is a three times winner of Loebners annual instantiation of Turings Test for machine intelligence [10].

When computer science evolves, so does the chatbot technology. In an aspect of managing knowledge-based data (some call it as the chatbots brain), an evolvement in chatbots architecture can be justified. The first chatbot ELIZA stored/embedded its data directly into the applications code, while more advanced A.L.I.C.E. later uses custom design language named Artificial Intelligence Markup Language or AIML (a derivative of Extensible Markup Language or XML) to manage its knowledge-based data [12],[14]. With more evolve Relational Database Model design together with Database Management System (DBMS) technology, came more advance chatbots. One of an example is VP-bot, a SQL-Based chatbot for medical application [7]. Developed by Dr. Weber from Harvard University, VPbot is a chatbot that takes advantage of Relational Database Model design to stored, manage and even uses the SQL language (database scripting language) to perform the chatbot keywords/pattern matching process.

2 Proposed Algorithms for Chatbots Issues

2.1 Extension and Prerequisite for Response's Relation

Extension and Prerequisite are proposed to enable relations between responses in chatbot technology. Designed by using an approach of Relational Database Model, Extension and Prerequisite was implemented both in keywords matching and knowledge-based authoring process. Typical chatbots are designed to response for users input in a one-way input-response paradigm without any parameter that holds the conversation issue. It was like a search engine where user typed an input and engine will produce an output based on that input alone. Later if a new search parameter is being entered, the search process will start again without any relation/reference to the previous search's issue. Therefore, in general chatbot process model (Fig. 1), input 1 will return a response 1, input 2 will return a response 2, and so on until the last input from user, an input n will return a response n.

Although there is a used of some techniques that will hold the topic of the conversation (AIML <topic> or <that> tags [14] and VPbot topic parameter [7]), those techniques does not exactly hold the conversation issue because the topic mechanism is a technique that replaces word/phrase with another word/phrase that had been stored as a constant variable at that particular conversation. There is also an argument that suggested the irrelevantly of the response given by AIML chatbot prior to the conversation issue. Jia stated that within two or more rounds, most users could find that the responses from chatbot are stupid

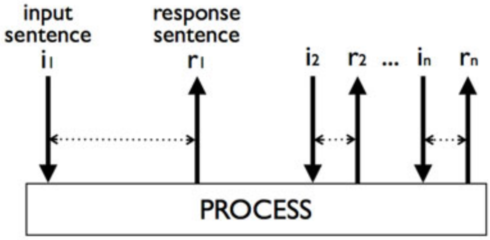


Fig. 1. General chatbot process model

and irrelevant with the topic and the context [3]. Shawar and Atwell stated that there is a logical inconsistencies in chatbot replies given by example that previous chatbot sentence suggested a non-enthusiasm about sport but later become enthuses about football [11].

In other words, topic is basically used as a replacement over pronoun with a constant noun. Example conversation implementing topic mechanism is; I broke my hand, Did it hurt?, replacement of pronoun it to the constant noun hand. Whereby, the real matter in holding conversation issue is a hypothetical conversation lines that human draw when they had a conversation with each other. For example, as human talk about car, What brand is your car? Toyota, How much is it? A thousand Dollar and later changed to the issue of house within the same conversation, Do you live in a bungalow? Yes, Where is it located? Near the beach. From this example, there is a hypothetical line regarding relations between responses in the issue of car and another line in the issue of house. This line is basically a connection that had been created in human conversation from responses that relate to each other. Extension and Prerequisite is attempt to create this line in a human conversation with chatbot so that in chatbot process model, it will become as in Fig. 2 where previous response is relate to current response and so on (P is Prerequisite data and E is Extension data).

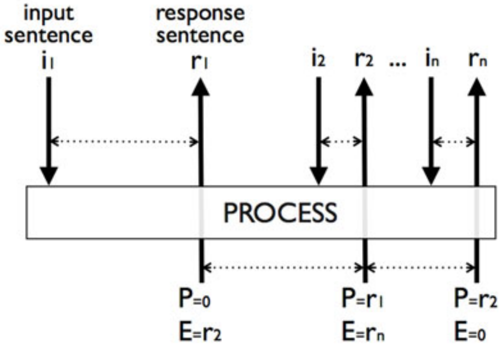


Fig. 2. Chatbot process model with response's relation

As proposed, tested and implemented by SQL-Based chatbot named ViDi (acronyms for Virtual Diabetes physician) [5], two components need to be incorporated into chatbot architecture in order to actualize Extension and Prerequisite algorithm into the matching process. Those components are; 1) unique ID for every response, 2) Extension and Prerequisite variable attach to every response that will used to hold the response's ID/s (Extension to hold next possible response/s while Prerequisite to hold previous possible response/s). Implementing Extension and Prerequisite, the algorithm steps regarding chatbot processes of receive input from user, performing keywords matching process and generate response back to user become as follows (for details on implementation of Extension and Prerequisite, please refer to [5]):

1. From current response, hold current response's Extension data (if any).
2. Receive another input from user.
3. Process input (normalization, synonyms replacement and so on).
4. Analyze Extension data gathered from Step 1.
5. Keywords matching regarding Extension ID/s (if Step 1 didn't hold any ID, proceed to Step 7).
6. If match, generate response and hold new responses Extension data (if any). If no match, proceed to Step 7.
7. Run keywords matching process towards the entire keywords database.
8. If match, generate response and hold new response's Extension data (if any).
9. If no match, generate response for user to enter another input. Hold same Extension data as in Step 1.

2.2 OMAMC for Matching Precedence and Flexibility

Reviewing ELIZA's keywords matching technique, an input sentence is analyzed from left to right. Each word is looked up in a dictionary of keywords for a match and if word/s is identified as keywords, then decomposition rule will apply [15] (note that decomposition rule is a method used by ELIZA in the process of reassembly rule or response generation). For A.L.I.C.E., its knowledge about English conversation is stored using a mechanism called Graphmaster (written using AIML). The Graphmaster consists of collection of nodes called Nodemappers. These Nodemappers map the branches from each node. The branches are either single words or wildcards. A convenient metaphor for AIML patterns is the file system stored in computers that are organized hierarchically (tree structure). The file system has a root, such as "c:/" and the root have some branches that are files, and some that are folders. The folders, in turn, have branches that are both folders and files. The leaf nodes of the whole tree structure are files. Every file has a "path name" that spells out its exact position within the tree. The Graphmaster is organized in exactly the same way. AIML that stored a pattern like "I LIKE TO *" is metaphorically are "g:/I/LIKE/TO/star". All of the other patterns that begin with "I" also go into the "g:/I/" folder. All of the patterns that begin with "I LIKE" go in the "g:/I/LIKE/" subfolder. So it's like the folder "g:/I/LIKE/TO/star" has a single file called "template.txt" that contains the template [12],[14].

From above literature, chatbots keywords/pattern matching techniques can be divided into two categories. First is rather similar to human brain incremental parsing technique where an input sentence is being analyzed in a word-by-word basis from left to right by sequence [8]. Keywords can be one-word keywords or many-words keywords but each word in many-words keywords must be attached to one another, forming a long keywords pattern (cannot be separated as e.g. one word in prefix and one word in suffix separated by several words in the middle). Second is a direct match process where input sentence is being analyzed for an appearance of keywords anywhere in the input sentence. Whole input sentence is being treated as a one variable and available keywords in the database will scan this variable for match.

The principal difference between first and second technique is first being input centered (words from input sentence is being matched against keywords in knowledge-based) and second being keywords centered (keywords in knowledge-based is being matched against an input sentence). Despite the difference, both categories suggested the same paradigm for matching process in which only one keywords is needed in order to trigger the respective response. One keywords in this context means one word, phrase or even sentence for one keywords set (not a collection of word, phrase or sentence). However, there is an augment regarding this matter by VPbot's keywords architectural design. In VPbot, author can assign several keywords (maximum of three) in the same keywords set. All keywords within the same set must be matched in order to trigger the respective response [7]. Using the second category of keywords matching technique, all keywords can be located anywhere in the input sentence and as long as the keywords is in the same set, VPbot will matched it. For the issue of precedence over which keywords is more accurate, longer keywords appear to have the top priority justified by long keywords set will only match a very specific phrase, while short keywords set will match a larger range of possible input queries [7].

One-match and All-match Categories or OMAMC technique comprises of two components that correlated with each other. The components are; 1) keywords arrangement (for matching precedence) and 2) keywords variety (for matching flexibility). Describing the fundamental idea of OMAMC, each response in ViDi's knowledge-based is designed to have an infinite number of keywords sets associated with either One-match or All-match category (note that OMAMC had been tested and implemented by the same SQL-Based chatbot named ViDi used by Extension and Prerequisite). Each keywords set in One-match category contains single keywords that can be in a form of one-word or many-words keywords (a single word or a phrase) while each keywords set in All-match category contains more than single keywords as in VPbot's keywords design. The different is that All-match keywords had no limit over how many keywords can a single set have (VPbot limitation is three keywords for each set). All-match keywords can be in a form of combination between a single word and a phrase, producing either multiple one-words keywords, multiple phrases keywords or combination of one-word/s and phrase/s keywords in the same single keywords set.

For both One-match and All-match categories, each keywords set will be stored as a single variable. Therefore, for All-match category that can have multiple keywords within the same set, author need to put a symbol of commas (",") to separate each keyword. For matching process, One-match is equalize to an exact-match process where word/s and its location must be the same as in the input sentence, while All-match is equalize to a flexible-match where words location is a flexible factor. Same as VPbot's keywords matching technique, if each All-match keywords within the same set is matched, the response will be triggered. The sequence location of the keywords can be different between the set and the input sentence. As example, first and second keywords in the set do not have to be in the same sequence location as in the input sentence (in the input sentence, the second keywords can come first before the first keywords).

Looking back to the two components of OMAMC (keywords arrangement for matching precedence and keywords variety for matching flexibility), keywords arrangement for this technique was designed based on keywords precedence as in literature, long keywords over short keywords (note that the length of keywords is defined by a total count of words within each set) and exact-match over flexible-match (one keywords over generic keywords) that is One-match over All-match. For keywords variety, OMAMC technique had expanse VPbot's technique on generic keywords by making no limitation on the number of keywords that can be associated with a single set. For details on implementation of OMAMC, please refer to [6].

3 Significance of Results and Contributions

Significance of results from implementing both proposed algorithms can be presented using following comparison tables (Table 1 and Table 2). Measurements are being done in regard to several focus area in which the discovered issues had been identified. Comparisons are being done towards both A.L.I.C.E.'s AIML and VPbot algorithms as being the two most referred chatbot in this investigation. Table 1 will present Extension and Prerequisite enhancement against chatbot's topic mechanism which are AIML <that> tag and VPbot topic parameter, while Table 2 will present OMAMC enhancement against AIML Graphmaster and VPbot's keywords set.

Significant contributions in regard to both proposed algorithms can be derived from presented comparison tables. Extension and Prerequisite algorithm have enabled the relations between responses in chatbot's chatting conversation. The relations created by this algorithm is a specific interaction between responses that relate to each other in the context of a whole sentence, not as the "topic" mechanism (by previous chatbot) that basically replaces word/phrase/sentence with another word/phrase/sentence that had been stored as a constant variable at each particular conversation. Although AIML <that> tag also support the context of a whole sentence, the implementation was rather unproductive considering the need to repetitively write the same template in every <that> tag for possible expected patterns.

Table 1. Extension and Prerequisite against topic mechanism

	Extension and Prerequisite	AIML Topic	VPbot Topic
Direct/exact relation/s between response/s	Yes	No	No
Matching precedence	Extension response/s, then Others (whole keywords database)	<that> tag, then Others (whole keywords database)	Whole keywords database (lack of true support of context)
Support for same keywords representing different meaning	Yes	Yes (AIML complexity, writing previous utterance for each pattern is a tedious activity)	No
Possible link for the whole conversation	Yes (draw a path)	No (did not draw a path)	No (did not draw a path)
Human Working Memory imitation (storing previous utterance)	Yes (can store the whole utterance)	Yes (can store the whole utterance)	Not particularly (cannot store the whole utterance)

Table 2. OMAMC against AIML Graphmaster and VPbot's keywords set

	OMAMC	AIML Graphmaster	VPbot Keywords set
Longer keywords effect on matching precedence	Yes	No	Yes
Precedence analysis while keywords matching	While matching	While matching	After matching
Exact-match precedence priority against other matching types	Highest (respective category)	Highest (same category)	Highest (same category)
Benchmark for stopping matching process if a match is found	Different category and/or lower words count	No benchmark (stop instantly)	No benchmark (matching all)
Generic keywords support	Yes (unlimited)	No	Yes (maximum of three)

Responses' relations have opened the possibility for chatbot to have a conversation that focus on one specific issue per time of the conversation. This scenario will create a more productive process for both chatbot (keywords matching) and users (input entering). For chatbot, implementing Extension and Prerequisite will reduce the processing time for keyword matching process. This is because algorithm have narrow down the size of keywords to be match from an entire keywords database, to just a keywords that correlate with the Extension data. As for users, input entering will become much easier when chatbot is capable of understanding elliptic syntax as an input. Elliptic input is an essential concern in chatbot processes because it was a general habit for human in chatting activity to response in such manner.

This issue is impartially related to a study in human cognitive function that suggested a Long-Term Working Memory component in human performing cognitive task. The example given is when human reading a sentence in a text, they must have access to previously mentioned actors and objects and also a contextual information (in that particular subject) in order to integrate coherently with the information presented in the current sentence [2]. This relation also supported the issue of the needs for direct instructional guidance rather than minimally-guided or unguided instruction that is less effective and efficient on human learning process [4]. In the context of chatbot that functions as knowledge representation system, relations between responses can be substantially used for the guided-chatting activity in which chatbot can keep giving guidance on how to proceed. This chatting scenario will principally eliminate the idle time when users did not know what to response and later leaving the chatbot as they become bored.

In OMAMC implementation, the first issue to be analyzed is precedence. For AIML with Graphmaster component, precedence for keywords goes by atomic categories (exact- match), then default categories (pattern with wildcard/s) and lastly recursive categories (symbolic reduction, synonyms replacement). To be noted that in AIML, longer keywords will not affect the precedence level. For VPbot, all keywords will be matched first before precedence analysis is being done. VPbot precedence goes by specific instance over generic response (exact-match over flexible-match), variation with low total weighs over high total weights (symbolic reduction, synonyms replacement) and lastly total string length (longer string over shorter string). For both techniques, exact-match is considered to be the highest precedence over all keywords.

In OMAMC, exact-match keywords is treated in a total different category from generic keywords (flexible-match) with One-match being the exact-match and All-match being the flexible-match. Being in different category, if algorithm finds a match in One-match category, then All-match category will not be processed. This will eliminate the redundant matching time for less precedence keywords if more precedence keywords had already been matched. Later if no match is found within One-match category, algorithm will proceed to generic keywords match that is an All-match category.

With strong argument by VPbot that longer string length have more precedence over short string length, One-match and All-match keywords had built in attached variable name ”wordCount” to encounter this issue. In each category according to precedence (One-match then All-match), wordCount will be among the first to be analyzed in order to avoid unnecessary matching process. That is if a match is found, wordCount for that keywords will be saved as a benchmark for string length. Therefore, algorithm will not process keywords with less count of words than already matched keywords, eliminating the need for unnecessary matching process for keywords that eventually will not be used.

The second issue to be analyzed is matching flexibility that is created by generic keywords technique. AIML did not have the support for generic keywords while VPbot had the limit of maximum three keywords for each set (keywords 1, 2 and 3). For All-match category, generic keywords had no limit in quota (keywords 1 to n). Same rule as VPbot is applied where all keywords within the same set must be matched in order to trigger the response.

4 Conclusion

In this paper, the significance of results and contributions from two proposed algorithms has been presented. The algorithm named Extension and Prerequisite had enabled chatbot to have relations between responses that open up a possibility for chatbot to have a specific issue conversation in a more controlled approach. This functionality also makes the chatting activity become more productive for human and chatbot itself prior to the focus issue being the main concern. As a result, telegraphic/elliptic inputs by users become understandable by chatbot and processing time regarding finding a keywords becoming much faster.

The second algorithm named OMAMC is proposed to enhance chatbot's keywords matching technique in the context of keywords arrangement for matching precedence and keywords variety for matching flexibility. Other area in which OMAMC technique can be implemented is in computer hardware processing algorithms that involved in string matching process [9]. In this area, further research can be done into making the two categories of OMAMC being process in two different string matching algorithms with One-match category being directly match without preprocessing phase, and All-match category being match with preprocessing phase (because the flexible matching process of generic keywords). Differentiating these two processes could result in 1) faster processing time by the reason that All-match category did not have to be matched if One-match category already found a match and 2) maintaining matching flexibility for generic keywords category (All-match category) while still concerning the processing time for exact match keywords category (One-match category). From interconnectivity between OMAMC and other areas of computing, it can be said that OMAMC technique is also and could be useful in many areas despite the original design purpose is for the used of keywords matching process in chatbot technology.

References

1. Christy, A., Thambidurai, P.: CTSS: A Tool for Efficient Information Extraction with Soft Matching Rules for Text Mining. J. Comput. Sci. 4, 375–381 (2008); doi:10.3844/.2008.375.381
2. Ericsson, K.A., Kintsch, W.: Long-term working memory. Psychol. Rev. 102(2), 211–245 (1995); doi:10.1.1.20.2523

3. Jia, J.: The study of the application of a web- based chatbot system on the teaching of foreign languages. In: Proceeding of the Society for Information Technology and Teacher Education International Conference (SITE 2004), pp. 1208–1211. AACE Press, USA (2004), http://www.editlib.org/p/13634
4. Kirschner, P.A., Sweller, J., Clark, R.E.: Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. Educational Psychologist 41(2), 75–86 (2006), doi:10.1207/s15326985ep4102_1
5. Lokman, A.S., Zain, J.M.: Extension and prerequisite: An algorithm to enable relations between responses in chatbot technology. Journal of Computer Science 6(10), 1212–1218 (2010), doi:10.3844/jcssp.2010.1212.1218
6. Lokman, A.S., Zain, J.M.: One-Match and All-Match Categories for Keywords Matching in Chatbot. American Journal of Applied Science 7(10), 1212–1218 (2010), doi:10.3844/ajassp.2010.1406.1411
7. Ohno-Machado, L., Weber, G.M.: Data Representation and Algorithms for Biomedical Informatics Applications. Harvard University, Cambridge, MA 184, 11988–11989 (2005) ; ISBN: 0-542- 11988-9
8. Pickering, M., Clifton, C., Crocker, M.W.: Architectures and Mechanisms in Sentence Comprehension, pp. 1–28. Cambridge University Press, Cambridge (2000); ISBN: 0-521-63121-1
9. Raju, S.V., Babu, A.V.: Parallel Algorithms for String Matching Problem on Single and Two Dimensional Reconfigurable Pipelined Bus Systems. J. Comput. Sci. 3, 754–759 (2007), doi:10.3844/.2007.754.759
10. Shah, H.: ALICE: An ACE in digital and. tripleC 4, 284–292 (2006) ;ISSN: 1726-670X
11. Shawar, B.A., Atwell, E.: A chatbot system as a tool to animate a corpus. ICAME J. 29, 5–24 (2005), http://icame.uib.no/ij29/ij29-page5-24.pdf
12. Shawar, B.A., Atwell, E.: Chatbots: Are they really useful. LDV-Forum Band 22, 31–50 (2007), doi:10.1.1.106.1099
13. Turing, A.M.: Computing Machinery and Intelligence. In: Epstein, R., et al. (eds.) Parsing the Turing Tes, pp. 23–65. Springer Science + Business Media BV, Heidelberg (2009); ISBN: 978-1-4020-9624-2
14. Wallace, R.S.: The Anatomy of ALICE. In: Epstein, R., et al. (eds.) Parsing the Turing Test, pp. 181–210. Springer Science + Business Media B.V, Heidelberg (2009); ISBN: 978-1-4020-9624-2
15. Weizenbaum, J.: ELIZA-a computer program for the study of natural language communication between man and machine. Commun. ACM. 9, 36–45 (1966), doi:10.1145/365153.365168

An Alternative Measure for Mining Weighted Least Association Rule and Its Framework

Zailani Abdullah[1], Tutut Herawan[2], and Mustafa Mat Deris[3]

[1] Department of Computer Science, Universiti Malaysia Terengganu
Gong Badak, 21030 Kuala Terengganu, Terengganu, Malaysia
zailania@umt.edu.my

[2] Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
tutut@ump.edu.my

[3] Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat 86400, Johor, Malaysia
mmustafa@uthm.edu.my

Abstract. Mining weighted based association rules has received a great attention and consider as one of the important area in data mining. Most of the items in transactional databases are not always carried with the same binary value. Some of them might associate with different level of important such as the profit margins, weights, etc. However, the study in this area is quite complex and thus required an appropriate scheme for rules detection. Therefore, this paper proposes a new measure called Weighted Support Association Rules (WSAR*) measure to discover the significant association rules and Weighted Least Association Rules (WELAR) framework. Experiment results shows that the significant association rules are successfully mined and the unimportant rules are easily differentiated. Our algorithm in WELAR framework also outperforms the benchmarked FP-Growth algorithm.

Keywords: Weighted, Association rules; Significant; Measure.

1 Introduction

Association rules mining is playing an important role in knowledge discovery for more than a decade. The main objectives of association rules mining are to search for the interesting correlations, associations or casual structures among sets of items in the data repositories. It was first initiated by Agrawal *et al.* [1] and amazingly still continuing as one of the active research areas. In association rules, a set of item is defined as an itemset. The itemset is said to be frequent, if it occurs more than a predefined minimum support. Besides that, confidence is another alternative measure used for association rules. The association rules are said to be strong if they appear more than a predefined minimum confidence. Least itemset is a set of item that is rarely found in the database but may produce a useful insight for certain domain applications. The least itemset occurs very rare in the database but surprisingly that they might co-occurs in a high proportional with other specific itemset. In fact, this itemset

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 480–494, 2011.

is very important to some applications regardless of its tiny of support value. For examples, it can be used in detecting the unexpected parameters in air pollution, network intruders, critical faulty [4], diseases symptoms [2], and many more. However, the traditional approachs rooted in the crisp and uniform minimal support are hardly in facilitating such uncommon itemsets [3]. By increasing the minimum support, many potential itemsets and rules are truncated out during the pruning process. Therefore, a simple approach is by decreasing the minimum support until the least itemsets are found. But, the trade off is it may generates a huge number of both significant and unnecessary of itemsets. Furthermore, the low minimum support will also proportionally increase the computational cost and its complexity. This unavoidable drawback is also known as the rare item problem [7]. Mining frequent itemsets or association rules are considered one of the most famous studies in data mining. Hence, there are quite a number in term of efficient methods and techniques have been introduced in the past [1, 3, 5, 7-10, 13, 15-17]. Here, all items in transactional database are assumed to have an equal weight or also known as binary weight (1 or 0). Indeed, the frequent itemset are always relied on two standard measures; support and confidence. However, this assumption is not always accurate since there are many cases that the items hold their own weight. In fact, the weight can be used to represent the important of the item in the transactional databases such as the price, profits margin, quantity, etc. Theoretically, mining weighted least itemset or association rules is not similar to mine binary frequent itemsets. The following scenario elaborates the limitation of mining tradition rules. For instance, in market basket analysis the manager wants to find out the association rules with the certain predefined conditions such as the item with the highest profit margins. Let assume that the profit of selling the smart phone is more than selling the cell phone accessories. Hence, the association between SIM card and smart phone is more significant than the association between SIM card and cell phone accessories. However, without considering the profit margin for each individual item, it is impossible to discover the most significant or interesting rules. Thus, the transactional items should be able to hold their own weight and a special measure should be derived or employed to trigger such association rules. On top of that, the performance issue is another challenge in mining these rules. Most of the previous studies are still incorporated with Apriori-like algorithms. As a result, these types of algorithms are suffered from two non-trivial costs [11]; generating of a huge number of candidate itemsets and repeatedly scanning the database to find the large itemsets. For k-itemsets, Apriori will produce up to 2k – 2 candidates in total. As a result, several studies have changed the strategy by employing the frequent pattern growth algorithm to mine the rules. It outperforms the typical Apriori-like algorithms. However, it may not fit into memory if the dataset size is very huge and the minimum support threshold is set to very low. Here, we have summarized three main contributions to solve the mentioned above problems. First, the Weighted Support Association Rules (WSAR*) measure is proposed to as a new measure for weighted association rules. The item support and its weight are utilized in formulating the WSAR*. Second, an enhanced version of existing prefix tree and frequent pattern growth algorithm called LP-Tree and LP-Growth algorithm is employed [14]. Hash-based approach [15] is employed to reduce the complexities and increase the computational performance of the algorithm. Third, experiments with benchmarked UCI dataset repositories [20] are performed to evaluate the scalability of the framework.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 explains the proposed method. Section 4 discusses the framework comparison. This is followed by comparison tests in section 5. Finally, conclusion and future direction are reported in section 6.

2 Related Work

Up to date, only few efforts have been made to discover the weighted association rules as compared to mine the typical binary rules. Three dominant factors of discouraging research in this field are computational cost, complexity and appropriate measures. There are several works have been carried out for the past decades in order to discover the significant itemset. Szathmary *et al.* [18] proposed Apriori-Rare model to extract rare itemsets with high confidence and low support from large databases. However, the main drawbacks of this model are the restoration of rare itemset is a very computational extensive and may generate a huge number of unnecessary rules. Cai *et al.* [12] introduced Weighted Association Rules (WAR) with MINWAL(O) and MINWAL(W) algorithms based on the support bounds approach to mine the weighted binary ARs. However, these algorithms are quite complicated and very time consuming. Selvi *et al.* [16] introduced Dynamic Collective Support Apriori (DCS-Apriori) to produce an interesting rare association rules by using two auto-counted minimum supports. However, the model is not yet tested using the real dataset and still suffers from candidate itemset generations. Kiran *et al.* [5] suggested an Improved Multiple Support Apriori Algorithm (IMSApriori) with support of different notion. However, it still cannot avoid from facing the problems of rule missing and rule explosion. Zhou *et al.* [6] proposed two approaches to generate the least association rules called Matrix-based Scheme (MBS) and Hash-based scheme (HBS). The main drawbacks of MBS and HBS are memory space consumption and expensive cost of collision for unlimited items length, respectively. Koh *et al.* [7] proposed a novel Apriori-Inverse algorithm to mine the least itemset without generating any frequent rules. However, the main challenges are it still suffers from too many candidate itemset generations and computational times during generating the least association rules. Yun *et al.* [8] introduced a Relative Support Apriori Algorithm (RSAA) toward generating the least itemset from database. The main constrain of this algorithm is it increases the computational cost if the minimum relative support is set close to zero. In addition, determination of three predefined measurements is also another issue for this algorithm. Liu *et al.* [9] suggested algorithm called Multiple Support Apriori (MSA) to capture the least association rules. However, if the predefined MIS, lowest item minimum support LS and values are set to very high or very low, this algorithm is still suffered from the "rare item problem". Wang *et al.* [10] proposed Adaptive Apriori to capture the required itemset. Several support constraints are used to each itemset. However, this algorithm still suffers from necessity of scanning multiple times of database for generating the required itemset.Tao *et al.* [11] proposed an algorithm namely Weighted Association Rule Mining (WARM) to discover significant weight of itemset. However, the structure of this algorithm is still resembles the Apriori algorithm and it is not suitable for data types without having a preassigned weights. Ding [13] suggested a Transactional Co-Occurrence Matrix (TCOM) model to mine the least association rules. TCOM structure utilized the advantage of transactional and item oriented layout of the database. Although the theory behind the model is very advance, but implementation wise is quite costly and impractical. In summary, a main basic concept underlying the proposed approaches [3-10] is still relying on the Apriori-like algorithm. The test-and-generate strategy is still the main concerns and open problems. If the varied minimum support threshold is set close to zero, these approaches will take similar amount of time as taken by Apriori. Most of previous approaches are required to set up a minimum support to be very low in order to capture the least items. As a result, enormous mixed of rules will be generated. Therefore, any approach to discover weighted association rules should try to evade from employing Apriori-like algorithms. However, implementation wise for others than tradition Apriori-like algorithm is not straight forward. Currently, FP-Growth [17] is considered as one of the fastest approach and benchmarked algorithm for frequent itemset mining. This algorithm can break two bottlenecks of Apriori-like algorithms. Yet, this algorithm is not scalable enough in mining the significant association rules and due to its limitation of static minimum support threshold.

3 Proposed Method

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items, $W = \{w_1, w_2, \cdots, w_{|A|}\}$, refers to the set of literals called set of weights with a non-negative real numbers, and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \leq |M| \leq |A|$ and each transaction can be identified by a distinct identifier TID.

3.1 Definition

Definition 1. (Least Items). *An itemset X is called least item if* $\alpha \leq \text{supp}(X) \leq \beta$, *where* α *and* β *is the lowest and highest support, respectively.*

The set of least item will be denoted as Least Items and

$$\text{Least Items} = \{X \subset I \mid \alpha \leq \text{supp}(X) \leq \beta\}$$

Example 2. Let $I = \{1,2,3,4,5,6\}$, $W = \{0.1,0.3,0.9,0.4,0.2,0.1\}$ and $T = \{\{1,2,3,4\}, \{1,2,5\}, \{1,5\}, \{1,6\}, \{2,3,5\}, \{2,4,5\}\}$. Thus, the Least Items for Interval Least Support (*ILSupp*) $[0.2,0.4]$ will capture only items 3 and 4, i.e.,

$$\text{Least Items} = \{3,4\}$$

Definition 3. (Frequent Items). *An itemset X is called frequent item if* $\text{supp}(X) > \beta$, *where* β *is the highest support.*

The set of frequent item will be denoted as Frequent Items and

$$\text{Frequent Items} = \{X \subset I \mid \text{supp}(X) > \beta\}$$

Example 4. From transaction T in Example 8, the Frequent Items for $\beta > 0.3$, will capture items 1, 2 and 3, i.e.,

$$\text{Frequent Items} = \{1,2,5\}$$

Definition 5. (Merge Least and Frequent Items). *An itemset X is called least frequent items if* $\text{supp}(X) \geq \alpha$, *where* α *is the lowest support.*

The set of merging least and frequent item will be denoted as LeastFrequent Items and

$$\text{LeastFrequent Items} = \{X \subset I \mid \text{supp}(X) \geq \alpha\}$$

LeastFrequent Items will be sorted in descending order and it is denoted as

$$\text{LeastFrequent Items}^{\text{desc}} = \left\{ \begin{array}{l} X_i \middle| \text{supp}(X_i) \geq \text{supp}(X_j),\ 1 \leq i, j \leq k,\ i \neq j, \\ k = |\text{LeastFrequent Items}|,\ x_i, x_j \subset \text{LeastFrequent Items} \end{array} \right\}$$

Example 6. From transaction T in Example 8, the Frequent Items for $\beta > 0.4$, will capture items 1, 2 and 5 i.e.,

$$\text{LeastFrequent Items} = \{1,2,3,4,5\}$$

These leastfrequent items are then sorted in descending order.

$$\text{LeastFrequent Items}^{\text{desc}} = \{1,2,5,3,4\}$$

Definition 7. (Ordered Items Transaction). *An ordered items transaction is a transaction which the items are sorted in descending order of its support and denoted as* t_i^{desc}, *where*

$$t_i^{\text{desc}} = \text{LeastFrequentItems}^{\text{desc}} \cap t_i, 1 \leq i \leq n, \left|t_i^{least}\right| > 0, \left|t_i^{frequent}\right| > 0 .$$

An ordered items transaction will be used in constructing the proposed model, so-called LP-Tree.

Definition 8. (Significant Least Data). *Significant least data is one which its occurrence less than the standard minimum support but appears together in high proportion with the certain data.*

Example 9. (In the case of significant and critical least items are not discovered). From transaction T in Example 7, based on *ILSupp*$[0.2,0.4]$ both items 3 and 4 have support of 0.33 which classify as Least Items. Itemset $\{2,3\}$ and $\{2,4\}$ have supports of 0.33, respectively, but in contrast, item 2 has support of 0.67 and classifies as Frequent Item. Therefore, using the existing methods, itemset $\{2,3\}$ and $\{2,4\}$ are not discovered because they do not satisfy the minimum support, β of 0.4.

Definition 10. (Item Weight). *A weight of an item is defined as a non negative real number and it denoted as*

$$\text{Item Weight} = \{X \subset I \mid 0 \leq weight(X) \leq 1\}$$

Definition 11. (Itemset Length). *A weight of an item is defined as a non negative real number and it denoted as*

$$\text{Itemset Length} = \{X \subset I \mid 0 \leq weight(X) \leq 1\}$$

Definition 12. (Weighted Support Association Rules). *A Weighted Support Association Rules (WSAR*) is a weight of itemset by formulating the combination of the support and weight of item, together with the total number of support in either of them.*

The value of Weighted Support Association Rules denoted as WSAR* and

$$\text{WSAR}*(I) = \frac{((\text{supp}(\text{A}) \times \text{weight(A)}) + (\text{supp}(\text{B}) \times \text{weight(B)}))}{(\text{supp}(\text{A}) + \text{supp}(\text{B}) - \text{supp}(\text{A} \Rightarrow \text{B}))}$$

WSAR* value is determined by multiplying the summation of items weight from both antecedent and consequence, with the support of the itemset.

Example 13. (Weighted Support Association Rules). From the transaction T in Example 10, based on $ILSupp[0.2, 0.3]$ and value of Minimum WSAR*, $\text{MinWSAR*} \geq 0.7$, the calculation of WSAR* for itemset $\{3,5\}$ and $\{3,7\}$ are as follows :

$$\text{WSAR}*(\{2,3\}) = \frac{\left(\left(\frac{4}{6} \times 0.3\right) + \left(\frac{2}{6} \times 0.9\right)\right)}{\left(\frac{4}{6} + \frac{2}{6} - \frac{2}{6}\right)} = 0.75$$

$$\text{WSAR}*(\{2,3\}) = \frac{\left(\left(\frac{4}{6} \times 0.3\right) + \left(\frac{2}{6} \times 0.4\right)\right)}{\left(\frac{4}{6} + \frac{2}{6} - \frac{2}{6}\right)} = 0.49$$

Therefore, itemset $\{3,5\}$ is considered as a significant association rule since its WSAR* more than Min-WSAR*.

3.2 Algorithm Development

Determine Interval Least Support. Let I is a non-empty set such that $I=\{i_1,i_2,\cdots,i_n\}$, and D is a database of transactions where each T is a set of items such that $T \subset I$. An itemset is a set of item. A k-itemset is an itemset that contains k items. From Definition 3, an itemset is said to be least if it has a support count within a range of α and β , respectively. In brevity, a least item is an itemset that satisfies the predefined Interval Least Support (*ILSupp*).

Construct LP-Tree. A Least Pattern Tree (LP-Tree) is a compressed representation of the least itemset. It is constructed by scanning the dataset of single transaction at a time and then mapping onto a new or existing path in the LP-Tree. Items that satisfy the ILSupp are only captured and used in constructing the LP-Tree.

Mining LP-Tree. Once the LP-Tree is fully constructed, the mining process will begin using bottom-up strategy. Hybrid 'Divide and conquer' method is employed to decompose the tasks of mining desired pattern. LP-Tree utilizes the strength of hash-based method during constructing itemset in descending order. Intersection technique from definition 4 is employed to increase the computational performance and reduce the complexity.

Construct Weighted Least Association Rules (WELAR). The rule is classified as weighted least association rules (WELAR) if it fulfilled two conditions. First, WSAR* of association rule must be greater than predefined minimum WSAR*. Second, the antecedent and consequence of association rule must be either Least Items or Frequent Items, respectively. The computation of WSAR* of each association rule is employed from Definition 12. Figure 1 shows a complete procedure to construct the WELAR algorithm.

```
WELAR Algorithm
Specify $WSI^{\min}$
for $(WI_a \in WeightedItem)$ do
    for $(WFI_i \in WI_a \bigcap FrequentItems)$ do
        for $(WLI_i \in WI_a \bigcap LeastItems)$ do
            Compute $WSI(WFI_i,WLI_i)$
            if $(WSI(WFI_i,WLI_i) > WSI^{\min})$ do
                Insert $WELAR(WFI_i,WLI_i)$
            end if
        end for loop
    end for loop
end for loop
```

Fig. 1. WELAR Algorithm

3.3 The Model/Framework

There are four major components involved in producing the significant weighted least association rules (WELAR). All these components are interrelated and the process flow is moving in one-way direction. A complete an overview framework of WELAR is shown in Fig. 2.

Scan Transactional Databases. This is the first stage of the model. The selected dataset are read or uploaded. The dataset is in a format of flat-file. Practicality, this format takes up much less space than the structure file. In the dataset, each transaction (record) is presented in a line and each item is separated by a single space. The main sources of dataset are obtained from UCI Machine Learning Repositories and research articles.

Generate Least Patterns. The first process is to convert the transactional data into LP-Tree structure by LP-Tree technique. During this process, Interval Threshold (Interval Least Support) is provided. The second process is to mine the least patterns from LP-Tree using LP-Growth technique. Any patterns that are failed to fulfil the Interval Least Support will be pruned out.

Assign Weight to Rules. The extracted patterns are then converted into association rules and assigned with WSAR* measure. Determination of item weight is based on the importance of the item in the transactions. Any rules that less than Minimum WSAR* will be excluded. The association rules will be than segregated according to predefine Interval Threshold (Interval Weighted Support).

Significant Weighted Least Association Rules. This is the final stage of the model. All association rules are now have their own weight. Ideally, the weight of the rules is a combination of classical item or itemset support and their weight (importance). The determinations of which rules are really significant or meaningful are then will be carried out by the domain expert.

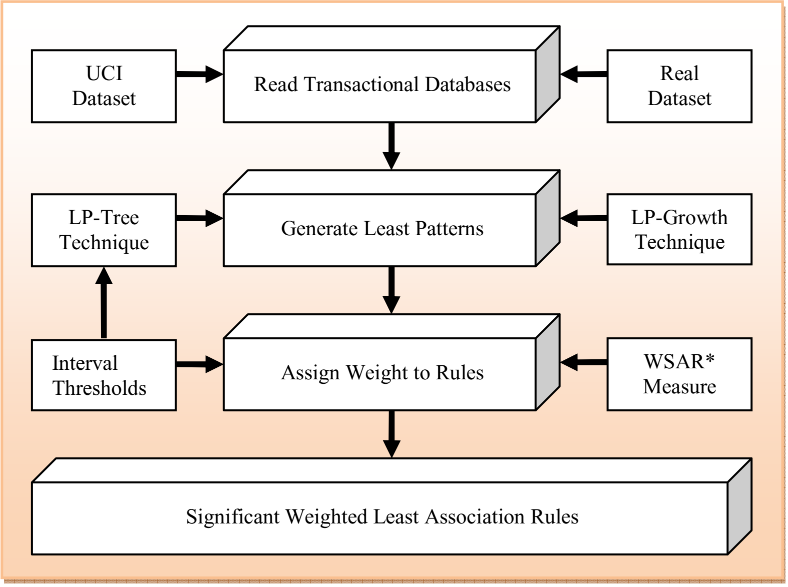


Fig. 2. An Overview Framework for WELAR

4 Framework Comparison

In this section, we do comparison tests between FP-Growth and LP-Growth algorithms. The performance analysis is made by comparing the processing time and number of iteration required. We used one simple dataset and two benchmarked datasets. These experiments have been conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. All algorithms have been developed using C# as a programming language.

4.1 A Dataset from [19]

In this section, we do comparative analysis of weighted least rules being generated using current weighted association rules, and the proposed measure, Weighted Support Association Rules (WSAR*). Table 1 shows 10 transactional databases with a maximum and minimum size of transaction are 5 and 1, respectively. Table 2 presents all items weight and its support. A range of interval supports used in mining least association rules for different measures are shown in Table 3.

Table 1. Transactional Database

TID	Items
T1	1 2 3
T2	2 4
T3	1 4
T4	3
T5	1 2 4 5
T6	1 2 3 4 5
T7	2 3 5
T8	4 5
T9	1 3 4
T10	2 3 4 5

Table 2. Items with weight and support

Items	Weight	Support
1	0.60	0.50
2	0.90	0.60
3	0.30	0.60
4	0.10	0.70
5	0.20	0.50

Table 3. Comparison of different weighted measures and their thresholds

Measures	Description	Interval Thresholds
CAR	Classical Association Rules (Agrawal *et al.,* 1993) [1]	10% - 30%
WAR	Weighted Association Rules (Cai *et al.,* 1993) [12]	0.1 – 0.3
WSSAR	Weighted Support Significant Association Rules (Tao *et al.,* 1993) [11]	0.1 – 0.3
WSAR*	Weighted Support Association Rules (proposed measure)	0.1 – 0.3

Table 4. Comparison of different weighted measures for least association rules

Rules	Support	WAR	WSSAR	WSAR*
3 1 --> 5	10%	0.11	0.36	0.35
4 2 --> 3	20%	0.26	0.43	0.46
2 3 --> 1	20%	0.36	0.60	0.68
4 3 --> 1	20%	0.2	0.33	0.34
4 2 --> 1	20%	0.32	0.53	0.56
2 1 --> 5	20%	0.34	0.56	0.58
1 --> 5	20%	0.16	0.40	0.50
4 1 --> 5	20%	0.11	0.45	0.31
4 3 --> 5	20%	0.18	0.20	0.21
4 --> 3	20%	0.12	0.20	0.25
2 --> 1	20%	0.45	0.75	1.05
3 --> 1	20%	0.27	0.45	0.60
3 --> 5	20%	0.15	0.25	0.35
4 2 --> 5	20%	0.36	0.40	0.47
2 3 --> 5	20%	0.42	0.47	0.55
4 --> 2	40%	0.4	0.50	0.68
2 --> 3	40%	0.48	0.60	0.90
4 --> 1	40%	0.28	0.35	0.46
2 --> 5	40%	0.44	0.55	0.91
4 --> 5	40%	0.12	0.15	0.21

Table 5. Comparison of different weighted measures and the occurrence of least association rules based on predefined interval thresholds

Measures	Total Rules
CAR	15
WAR	11
WSSAR	4
WSAR*	3

In overall, the total number of least association rules extracted using different measures is not similar. As in Tables 4 and 5, the result shows the lowest number of least association rules are captured by WSAR* and the highest are produced by CAR. From the 20 rules and based on the proposed measure, only 15% rules are classified as significant least association rules.

5 Comparison Tests

In this section, we do comparison tests between FP-Growth and LP-Growth algorithms. The performance analysis is made by comparing the number of rules extracted and the processing time required using variety of measures. The items weights are assigned randomly in a range of 0.1 to 1.0. For rules generation, only binary association between antecedent and consequence are taken into account. These experiments have been conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. All algorithms have been developed in .NET environment and C# as a programming language.

5.1 Retail Dataset from [20]

The first benchmarked dataset is Retails from Frequent Itemset Mining Dataset Repository. This dataset contains the retails market basket data from an anonymous Belgian retail store. Table 6 shows the fundamental characteristics of the dataset. The mapping between interval support and weighted interval support is presented in Table 7.

The number of significant least association rules being extracted from different types of measures with variety of interval thresholds is depicted in Fig. 3. From the total 2,404 of rules, WSAR* classified only 13% of them are significant least association rules. As compared to the other 3 measures, it is the lowest percentage. Therefore, our proposed measure can be considered as a good alternative to help in drilling down and finally determining the actuality of significance least association rules.

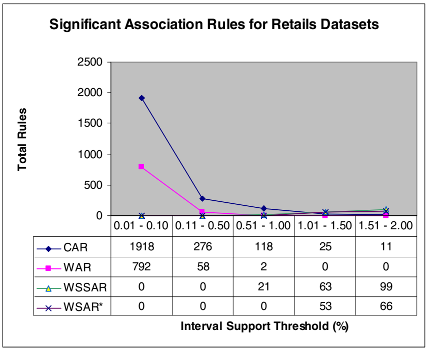
Fig. 4 shows the actual performance of both algorithms. In average, time taken for mining pattern sets for LP-Growth was 1.51 times faster than FP-Growth. Thus, this model is more scalable as compared to benchmarked FP-Growth. Generally, the processing time is decreasing once the *MinSupp* is increasing. This fact is applicable for both FP-Growth and LP-Growth.

Table 6. Retails Characteristics

Data sets	Size	#Trans	#Items	Average length
Retails	4.153 MB	88,136	16,471	10

Table 7. Mapping of interval support and interval weighted supports for Retails dataset

Interval Support (CAR)	Interval Weighted Supports (WAR, WSSAR, WSAR*)
0.01 - 0.10	0.001 - 0.010
0.11 - 0.50	0.011 - 0.050
0.51 - 1.00	0.051- 0.100
1.01 - 1.50	0.101 - 0.150
1.51 - 2.00	0.151 - 0.200



	0.01 - 0.10	0.11 - 0.50	0.51 - 1.00	1.01 - 1.50	1.51 - 2.00
CAR	1918	276	118	25	11
WAR	792	58	2	0	0
WSSAR	0	0	21	63	99
WSAR*	0	0	0	53	66

Fig. 3. Total significant least association rules generated from Retails dataset using different measures

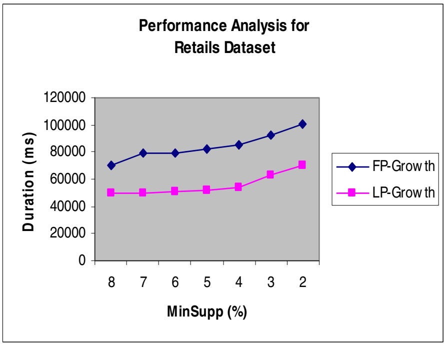


Fig. 4. Computational performance of mining the Retail dataset between FP-Growth and LP-Growth

5.2 Mushroom Dataset From [20]

The second and last benchmarked dataset is Mushroom from Frequent Itemset Mining Dataset Repository. It is a dense dataset and consists of 23 species of gilled mushroom in the *Agaricus* and *Lepiota Family*. Table 8 shows the fundamental characteristics of the dataset. The mapping between interval support and weighted interval support is presented in Table 9. For rules generation, as similar to above experiment, only binary association between antecedent and consequence are considered.

The total number of least association rules extracted using different types of measures and interval thresholds is presented in Fig. 5. From out of 2,871 of rules, only 11% of them are categorized by WSAR* as the significant least association rules. As compared to the other 3 measures, this is the lower percentage. Therefore and based on the previous experiment as well, our proposed measure is still the best option for zooming in details the most significance of the least association rules among them.

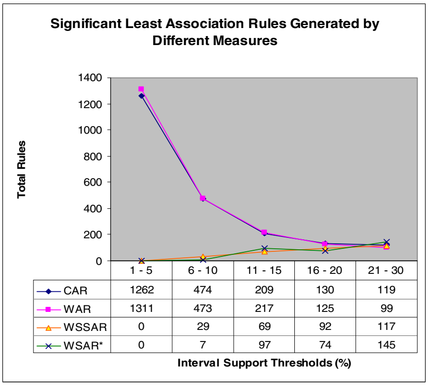
Fig. 6 illustrates the actual performance of both algorithms. In average, time taken for mining pattern sets for LP-Growth was 1.48 times faster than FP-Growth. Thus, this model is still scalable as compared to benchmarked FP-Growth. As similar to previous experiment, the processing time is decreasing once the *MinSupp* is increasing. In fact, this condition is also applicable for both FP-Growth and LP-Growth.

Table 8. Mushroom Characteristics

Data sets	Size	#Trans	#Items	Average length
Mushroom	0.83MB	8,124	120	23

Table 9. Mapping of interval support and interval weighted supports for Mushroom dataset

Interval Support (CAR)	Interval Weighted Supports (WAR, WSSAR, WSAR*)
1 - 5	0.01 - 0.05
6 - 10	0.06 - 0.10
11 - 15	0.11- 0.15
16 - 20	0.16 - 0.20
21 - 30	0.21 - 0.25



	1 - 5	6 - 10	11 - 15	16 - 20	21 - 30
CAR	1262	474	209	130	119
WAR	1311	473	217	125	99
WSSAR	0	29	69	92	117
WSAR*	0	7	97	74	145

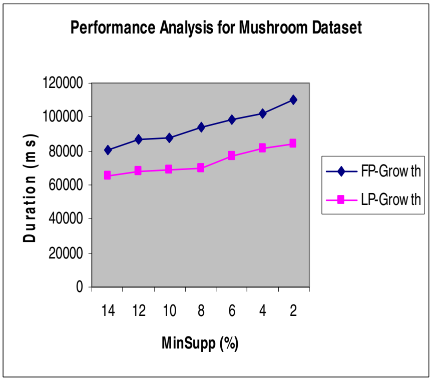


Fig. 5. Total significant least association rules generated from Mushroom using different measures

Fig. 6. Computational performance of mining the Mushroom dataset between FP-Growth and LP-Growth

6 Conclusion

Mining the weighted association rules is a very important study since not all items in transactional databases are carrying the same binary value. There are many situations that the item itself is associated with different levels of importance such as the profit margins, special offers, weights, etc. However, the complexity level of this research is very high as compared to mine the classical binary frequent rules. Moreover, a special measures and scalable framework are also become a crucial to deal with this problem. Therefore, in this paper we proposed a new weighted measure called Weighted Support Association Rules (WSAR*) and a novel Weighted Association Rules (WELAR) framework. We compared our proposed WSAR* with Classical Association Rules [1], Weighted Association Rules [12] and Weighted Support Significant Association Rule [11] in term of number of significant rules being extracted. Two datasets from Frequent Itemset Mining Dataset Repository [20] were used in the experiments. We do compare our algorithm in WELAR framework with the benchmarked FP-Growth algorithm. The result shows that our proposed measure and algorithm are outperformed the existing benchmarked measures and algorithm. It can discover the significant and least association rules from the large databases with an excellent computational performance.

In the future, we plan to evaluate our proposed solution into several real and benchmarked datasets.

References

1. Agrawal, R., Imielinski, T., Swami, A.,, D.M.: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 5(6), 914–925 (1993)
2. Lan, G.-C., Hong, T.-P., Tseng, V.S.: A Novel Algorithm for Mining Rare-Utility Itemsets in a Multi-Database Environment. In: The 26th Workshop on Combinatorial Mathematics and Computation Theory, pp. 293–302 (2009)
3. Weiss, G.M.: Mining with Rarity: a Unifying Framework. SIGKDD Explorations Newsletter 6(1), 7–19 (2004)
4. Abdullah, Z., Herawan, T., Deris, M.M.: Mining Significant Least Association Rules Using Fast SLP-Growth Algorithm. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN 2010. LNCS, vol. 6059, pp. 324–336. Springer, Heidelberg (2010)
5. Kiran, R.U., Reddy, P.K.: An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. In: Proceeding of IEEE Symposium on Computational Intelligence and Data Mining, pp. 340–347 (2009)
6. Zhou, L., Yau, S.: Assocation Rule and Quantative Association Rule Mining among Infrequent Items. In: Proceeding of ACM SIGKDD, Article No. 9 (2007)
7. Koh, Y.S., Rountree, N.: Finding Sporadic Rules Using Apriori-Inverse. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 97–106. Springer, Heidelberg (2005)
8. Yun, H., Ha, D., Hwang, B., Ryu, K.H.: Mining Association Rules on Significant Rare Data using Relative Support. The Journal of Systems and Software 67(3), 181–190 (2003)
9. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In: Proceeding of ACM SIGKDD 1999, pp. 337–341 (1999)

10. Wang, K., Hee, Y., Han, J.: Pushing Support Constraints into Association Rules Mining. IEEE Transactions on Knowledge and Data Engineering 15(3), 642–658 (2003)
11. Tao, F., Murtagh, F., Farid, M.: Weighted Association Rule Mining using Weighted Support and Significant Framework. In: Proceeding of ACM SIGKDD 2003, pp. 661–666 (2003)
12. Cai, C.H., Fu, A.W.C., Cheng, C.H., Kwong, W.W.: Mining Association Rules with Weighted Items. In: Proceedings of the international Database Engineering and Application Symposium, Cardiff, UK, pp. 68–77 (1998)
13. Ding, J.: Efficient Association Rule Mining among Infrequent Items. Ph.D. Thesis, University of Illinois at Chicago (2005)
14. Abdullah, Z., Herawan, T., Deris, M.M.: Scalable Model for Mining Critical Least Association Rules. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. LNCS, vol. 6377, pp. 509–516. Springer, Heidelberg (2010)
15. Park, J.S., Chen, M.-S., Yu, P.S.: An Effective Hash based Algorithm for Mining Association Rules. In: Carey, M.J., Schneider, D.A. (eds.) Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, pp. 175–186 (1995)
16. Selvi, C.S.K., Tamilarasi, A.: Mining association rules with dynamic and collective support thresholds. International Journal on Open Problems Computational Mathematics 2(3), 427–438 (2009)
17. Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns without Candidate Generation. In: Proceeding of the 2000 ACM SIGMOD, pp. 1–12 (2000)
18. Szathmary, L.: Generating Rare Association Rules Using the Minimal Rare Itemsets Family. International Journal of Software and Informatics 4(3), 219–238 (2010)
19. Khan, M.S., Muyeba, M., Coenen, F.: Weighted Association Rule Mining from Binary and Fuzzy Data. In: Proceedings of the 8th industrial conference on Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, Leipzig, Germany, July 16-18, vol. 212, pp. 200–212 (2008)
20. http://fimi.cs.helsinki.fi/data/

Mining Interesting Association Rules of Student Suffering Mathematics Anxiety

Tutut Herawan[1], Prima Vitasari[2], and Zailani Abdullah[3]

[1] Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang
Lebuhraya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia
tutut@ump.edu.my
[2] Centre of Modern Language and Human Sciences
Universiti Malaysia Pahang
Lebuhraya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia
primavitasari@yahoo.com
[3] Department of Computer Science, Faculty of Science and Technology
Universiti Malaysia Terengganu
Gong Badak 21030, Kuala Terengganu, Terengganu, Malaysia
zailania@umt.edu.my

Abstract. Up to this moment, association rules mining are one of the most important issues in data mining application. One of the commonly and popular techniques used in data mining application is association rules mining. The purpose of this study is to apply an enhanced association rules mining method, so called SLP-Growth (Significant Least Pattern Growth) proposed by [9] for capturing interesting rules in student suffering mathematics anxiety dataset. The dataset was taken from a survey on exploring mathematics anxiety among engineering students in Universiti Malaysia Pahang (UMP). The results of this research will provide useful information for educators to make a decision on their students more accurately, and to adapt their teaching strategies accordingly. It also can be helpful to assist students in handling their fear of mathematics and useful in increasing the quality of learning.

Keywords: Least association rules; Efficient; Critical least support, Mathematics anxiety.

1 Introduction

Anxiety is a psychological and physical response to treat a self-concept characterized by subjective, consciously perceived feelings of tension [1]. Anxious students have experience of cognitive deficits like misapprehension of information or blocking of memory and recall. Anxiety response to mathematics is a significant concern in terms of the perception that high anxiety will relate to avoidance of math tasks [2]. Mathematics anxiety is one of the psychological barriers that students encounter when they are performing a mathematics task [3]. Many mathematics educators find themselves overwhelmed with data, but lack the information they need to make informed

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 495–508, 2011.

decisions. Currently, there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community [4]. One of the popular data mining methods is Association Rules Mining (ARM). Association rules mining has been widely studied in knowledge discovery community [5]. It aims at discovering the interesting correlations, frequent patterns, associations or casual structures among sets of items in the data repositories. The problem of association rules mining was first introduced by Agrawal for market-basket analysis [6,7,8]. There are two main stages involved before producing the association rules. First, find all frequent items from transactional database. Second, generate the common association rules from the frequent items.

Generally, an item is said to be frequent if it appears more than a minimum support threshold. These frequent items are then used to produce the ARs. Besides that, confidence is another measure that always used in pair with the minimum support threshold. By definition, least item is an itemset whose rarely found in the database but it may produce interesting and useful ARs. These type of rules are very meaningful in discovering rarely occurring but significantly important, such as air pollution detection, critical fault detections, network intrusions, etc. and their possible causes. For the past developments, many series of ARs mining algorithms are using the minimum supports-confidence framework to avoid the overloaded of ARs. The challenge is, by increasing or decreasing the minimum support or confidence values, the interesting rules might be missing or untraceable. Since the complexity of study, difficulties in algorithms [2] and it may require excessive computational cost, there are very limited attentions have been paid to discover the highly correlated least ARs. In term of relationship, both of frequent and least ARs have a different degree of correlation. Highly correlated least ARs are referred to the itemsets that its frequency does not satisfy a minimum support but are very highly correlated. ARs are classified as highly correlated if it is positive correlation and in the same time fulfils a minimum degree of predefined correlation. Until this moment, statistical correlation technique has been successfully applied in the transaction databases [3], which to find relationship among pairs of items whether they are highly positive or negative correlated. As a matter of fact, it is not absolute true that the frequent items have a positive correlation as compared to the least items. In previous papers, we address the problem of mining least ARs with the objectives of discovering significant least ARs but surprisingly highly correlated [9,10]. A new algorithm named Significant Least Pattern Growth (SLP-Growth) to extract these ARs is proposed [9]. The proposed algorithm imposes interval support to extract all least itemsets family first before continuing to construct a significant least pattern tree (SLP-Tree). The correlation technique for finding relationship between itemset is also embedded to this algorithm [9]. In this paper, we explore SLP-Growth algorithm for capturing interesting rules in student suffering mathematics anxiety dataset. The dataset was taken from a survey on exploring mathematics anxiety among engineering students in Universiti Malaysia Pahang (UMP). The results of this research will provide useful information for educators to make a decision on their students more accurately, and to adapt their teaching strategies accordingly. It also can be helpful to assist students in handling their fear of mathematics and useful increasing the quality of learning.

The reminder of this paper is organized as follows. Section 2 describes the related work. Section 3 describes the basic concepts and terminology of ARs mining. Section 4 describes the proposed method, SLP-Growth algorithm. This is followed by performance analysis through mathematics anxiety dataset in section 5 and the results are presented in Section 6. Finally, conclusions of this work are reported in section 7.

2 Related Works

For the past decades, there are several efforts has been made to discover the scalable and efficient methods for mining frequent ARs. However, mining least ARs is still left behind. As a result, ARs that are rarely found in the database are pruned out by the minimum support-confidence threshold. As a matter of fact, the rarely ARs can also reveal the useful information for detecting the highly critical and exceptional situations. Zhou *et al.* [12] suggested a method to mine the ARs by considering only infrequent itemset. The drawback is, Matrix-based Scheme (MBS) and Hash-based scheme (HBS) algorithms are very expensive in term of hash collision. Ding [5] proposed Transactional Co-occurrence Matrix (TCOM for mining association rule among rare items. However, the implementation wise is quite complex and costly. Yun *et al.* [11] introduced the Relative Support Apriori Algorithm (RSAA) to generate rare itemsets. The challenge is, it takes similar time taken as performed by Apriori if the allowable minimum support is set to very low. Koh *et al.* [14] suggested Apriori-Inverse algorithm to mine infrequent itemsets without generating any frequent rules. However, it suffers from candidate itemset generations and costly in generating the rare ARs. Liu *et al.* [15] proposed Multiple Support Apriori (MSApriori) algorithm to extract the rare ARs. In actual implementation, this algorithm is facing the “rare item problem”. From the proposed approaches [11,12–15], many of them are using the percentage-based approach to improve the performance as faces by the single minimum support based approaches. In term of measurements, Brin *et al.* [16] introduced objective measure called lift and chi-square as correlation measure for ARs. Lift compares the frequency of pattern against a baseline frequency computed under statistical independence assumption. Omiecinski [17] proposed two interesting measures based on downward closure property called all confidence and bond. Lee *et al.* [18] suggested two algorithms for mining all confidence and bond correlation patterns by extending the pattern-growth methodology Han *et al.* [19]. In term of mining algorithms, Agrawal *et al.* [6] proposed the first ARs mining algorithm called Apriori. The main bottleneck of Apriori is, it requires multiple scanning of transaction database and also generates huge number of candidate itemsets. Han *et al.* [20] suggested FP-Growth algorithm which amazingly can break the two limitations as faced by Apriori series algorithms. Currently, FP-Growth is one of the fastest approach and most benchmarked algorithms for frequent itemsets mining. It is derived based on a prefix tree representation of database transactions (called an FP-tree).

3 Essential Rudiments

3.1 Association Rules (ARs)

ARs were first proposed for market basket analysis to study customer purchasing patterns in retail stores [6]. Recently, ARs has been used in many applications or disciplines such as customer relationship management [21], image processing [22], mining air pollution data [24]. Typically, association rule mining is the process of discovering associations or correlation among itemsets in transaction databases, relational databases and data warehouses. There are two subtasks involved in ARs mining: generate frequent itemsets that satisfy the minimum support threshold and generate strong rules from the frequent itemsets.

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \le |M| \le |A|$ and each transaction can be identified by a distinct identifier TID.

Definition 1. *A set* $X \subseteq I$ *is called an itemset. An itemset with k-items is called a k-itemset.*

Definition 2. *The support of an itemset* $X \subseteq I$, *denoted* $\mathrm{supp}(X)$ *is defined as a number of transactions contain X.*

Definition 3. *Let* $X, Y \subseteq I$ *be itemset. An association rule between sets X and Y is an implication of the form* $X \Rightarrow Y$, *where* $X \bigcap Y = \phi$. *The sets X and Y are called antecedent and consequent, respectively.*

Definition 4. *The support for an association rule* $X \Rightarrow Y$, *denoted* $\mathrm{supp}(X \Rightarrow Y)$, *is defined as a number of transactions in D contain* $X \bigcup Y$.

Definition 5. *The confidence for an association rule* $X \Rightarrow Y$, *denoted* $\mathrm{conf}(X \Rightarrow Y)$ *is defined as a ratio of the numbers of transactions in D contain* $X \bigcup Y$ *to the number of transactions in D contain X. Thus*

$$conf(X \Rightarrow Y) = \frac{\sup \mathrm{p}(X \Rightarrow Y)}{\sup \mathrm{p}(X)}.$$

An item is a set of items. A *k*-itemset is an itemset that contains *k* items. An itemset is said to be frequent if the support count satisfies a minimum support count (minsupp). The set of frequent itemsets is denoted as L_k. The support of the ARs is the ratio of transaction in D that contain both X and Y (or $X \bigcup Y$). The support is also can be considered as probability $P(X \bigcup Y)$. The confidence of the ARs is the ratio of transactions in D contains X that also contains Y. The confidence also can be considered as conditional probability $P(Y|X)$. ARs that satisfy the minimum support and confidence thresholds are said to be strong.

3.2 Correlation Analysis

After the introduction of ARs, many researches including Brin *et al.* [16] had realized the limitation of the confidence-support framework. Utilizing this framework alone is quite impossible to discover the interesting ARs. Therefore, the correlation measure can be used as complimentary measure together with this framework. This leads to correlation rules as

$$A \Rightarrow B \quad (\text{supp}, \text{conf}, \text{corr}) \tag{1}$$

The correlation rule is a measure based on the minimum support, minimum confidence and correlation between itemsets A and B. There are many correlation measures applicable for ARs. One of the simplest correlation measures is Lift. The occurrence of itemset A is independence of the occurrence of itemset B if $P(A \bigcup B) = P(A)P(B)$; otherwise itemset A and B are dependence and correlated. The lift between occurrence of itemset A and B can be defined as:

$$\text{lift}(A, B) = \frac{P(A \bigcap B)}{P(A)P(B)} \tag{2}$$

The equation of (4) can be derived to produce the following definition:

$$\text{lift}(A, B) = \frac{P(B \mid A)}{P(B)} \tag{3}$$

or

$$\text{lift}(A, B) = \frac{\text{conf}(A \Rightarrow B)}{\text{supp}(B)} \tag{4}$$

The strength of correlation is measure from the lift value. If $\text{lift}(A, B) = 1$ or $P(B \mid A) = P(B)$ $(\text{or } P(A \mid B) = P(B))$ then B and A are independent and there is no correlation between them. If $\text{lift}(A, B) > 1$ or $P(B \mid A) > P(B)$ $(\text{or } P(A \mid B) > P(B))$, then A and B are positively correlated, meaning the occurrence of one implies the occurrence of the other. If $\text{lift}(A, B) < 1$ or $P(B \mid A) < P(B)$ $(\text{or } P(A \mid B) < P(B))$, then A and B are negatively correlated, meaning the occurrence of one discourage the occurrence of the other. Since lift measure is not down-ward closed, it definitely will not suffer from the least item problem. Thus, least itemsets with low counts which per chance occur a few times (or only once) together can produce enormous lift values.

3.3 FP-Growth

Candidate set generation and tests are two major drawbacks in Apriori-like algorithms. Therefore, to deal with this problem, a new data structure called frequent pattern tree (FP-Tree) was introduced. FP-Growth was then developed based on this data structure and currently is a benchmarked and fastest algorithm in mining frequent itemset [19]. The advantages of FP-Growth are, it requires two times of scanning the transaction database. Firstly, it scans the database to compute a list of frequent items sorted by descending order and eliminates rare items. Secondly, it scans to compress the database into a FP-Tree structure and mines the FP-Tree recursively to build its conditional FP-Tree.

A simulation data [23] is shown in Table 1. Firstly, the algorithm sorts the items in transaction database with infrequent items are removed. Let say a minimum support is set to 3, therefore alphabets `f, c, a, b, m, p` are only kept. The algorithm scans the entire transactions start from T1 until T5. In T1, it prunes from {`f, a, c, d, g, i, m, p`} to {`f, c, a, m, p, g`}. Then, the algorithm compresses this transaction into prefix tree which f becomes the root. Each path on the tree represents a set of transaction with the same prefix. This process will execute recursively until the end of transaction. Once the complete tree has been built, then the next pattern mining can be easily performed.

Table 1. A Simple Data

TID	Items
T1	a c m f p
T3	b f h j o
T4	b c k s p
T5	a f c e l p m n

4 The Proposed Method

4.1 Algorithm Development

Determine Interval Support for least Itemset

Let *I* is a non-empty set such that $I = \{i_1, i_2, \cdots, i_n\}$, and *D* is a database of transactions where each *T* is a set of items such that $T \subset I$. An item is a set of items. A *k*-itemset is an itemset that contains *k* items. An itemset is said to be least if the support count satisfies in a range of threshold values called Interval Support (ISupp). The Interval Support is a form of ISupp (ISMin, ISMax) where ISMin is a minimum and ISMax is a maximum values respectively, such that $\text{ISMin} \geq \phi$, $\text{ISMax} > \phi$ and $\text{ISMin} \leq \text{ISMax}$. The set is denoted as L_k. Itemsets are said to be significant least if they satisfy two conditions. First, support counts for all items in the itemset must greater ISMin. Second, those itemset must consist at least one of the least items. In brevity, the significant least itemset is a union between least items and frequent items, and the existence of intersection between them.

Construct Significant Least Pattern Tree

A Significant Least Pattern Tree (SLP-Tree) is a compressed representation of significant least itemsets. This trie data structure is constructed by scanning the dataset of single transaction at a time and then mapping onto path in the SLP-Tree. In the SLP-Tree construction, the algorithm constructs a SLP-Tree from the database. The SLP-Tree is built only with the items that satisfy the ISupp. In the first step, the algorithm scans all transactions to determine a list of least items, LItems and frequent items, FItems (least frequent item, LFItems). In the second step, all transactions are sorted in descending order and mapping against the LFItems. It is a must in the transactions to consist at least one of the least items. Otherwise, the transactions are disregard. In the final step, a transaction is transformed into a new path or mapped into the existing path. This final step is continuing until end of the transactions. The problem of existing FP-Tree are it may not fit into the memory and expensive to build. FP-Tree must be built completely from the entire transactions before calculating the support of each item. Therefore, SLP-Tree is an alternative and more practical to overcome these limitations.

Generate Least Pattern Growth (*LP-Growth*)

SLP-Growth is an algorithm that generates significant least itemsets from the SLP-Tree by exploring the tree based on a bottom-up strategy. ‘Divide and conquer’ method is used to decompose task into a smaller unit for mining desired patterns in conditional databases, which can optimize the searching space. The algorithm will extract the prefix path sub-trees ending with any least item. In each of prefix path sub-tree, the algorithm will recursively execute to extract all frequent itemsets and finally built a conditional SLP-Tree. A list of least itemsets is then produced based on the suffix sequence and also sequence in which they are found. The pruning processes in SLP-Growth are faster than FP-Growth since most of the unwanted patterns are already cutting-off during constructing the SLP-Tree data structure. The complete SLP-Growth algorithm is shown in Figure 1.

4.2 Weight Assignment

Apply Correlation

The weighted ARs (ARs value) are derived from the formula (4). This correlation formula is also known by lift. The processes of generating weighted ARs are taken place after all patterns and ARs are completely produced.

Discovery Highly Correlated Least ARs

From the list of weighted ARs, the algorithm will begin to scan all of them. However, only those weighted ARs with correlation value that more than one are captured and considered as highly correlated. For ARs with the correlation less than one will be pruned and classified as low correlation.

```
Read dataset, D
Set Interval Support (ISMin, ISMax)
for items, I in transaction, T do
    Determine support count, ItemSupp
end for loop
Sort ItemSupp in descending order, ItemSuppDesc
for ItemSuppDesc do
    Generate List of frequent items, FItems > ISMax
end for loop
for ItemSuppDesc do
    Generate List of least items,  ISMin <= LItems < ISMax
end for loop
Construct Frequent and Least Items, FLItems = FItems U LItems
for all transactions,T do
if (LItems $\cap$ I in T > 0) then
if (Items in T = FLItems) then
  Construct items in transaction in descending order, TItemsDesc
          end if
     end if
end for loop
for TItemsDesc do
    Construct SLP-Tree
end for loop
for all prefix SLP-Tree do
    Construct Conditional Items, CondItems
end for loop
for all CondItems do
    Construct Conditional SLP-Tree
end for loop
for all Conditional SLP-Tree do
    Construct Association Rules, AR
end for loop
for all AR do
    Calculate Support and Confidence
    Apply Correlation
end for loop
```

Fig. 1. SLP-Growth Algorithm

5 Scenario on Capturing Rules

5.1 Dataset

The dataset was taken from a survey on exploring mathematics anxiety among engineering students in Universiti Malaysia Pahang (UMP) [3]. A total 770 students participated in this survey. The respondents were 770 students, consisting of 394 males and 376 females. The respondents are undergraduate students from five engineering faculties at Universiti Malaysia Pahang, i.e., 216 students from Faculty of Chemical and Natural Resources Engineering (FCNRE), 105 students from Faculty of Electrical and Electronic Engineering (FEEE), 226 students from Faculty of Mechanical Engineering (FME), 178 students from Faculty of Civil Engineering and Earth Resources (FCEER), and 45 students from Faculty of Manufacturing Engineering and Technology Management (FMETM).The survey's finding indicated that mathematics anxiety among engineering students are manifested into five dimensions, namely; (a)

Feel mathematic is difficult subject, (b) Always fail in mathematic, (c) Always writing down while mathematic class, (d) Anxious if don't understand, and (e) Lose interest of mathematic. To this, we have a dataset comprises the number of transactions (student) is 770 and the number of items (attributes) is 5 (refers to Table 2).

Table 2. Mathematics anxiety dataset

Dataset	Size	#Transactions	# Items
Mathematic anxiety	14KB	770	25

5.2 Design

The design for capturing interesting rules on in student suffering mathematics anxiety dataset is described in the following figure.

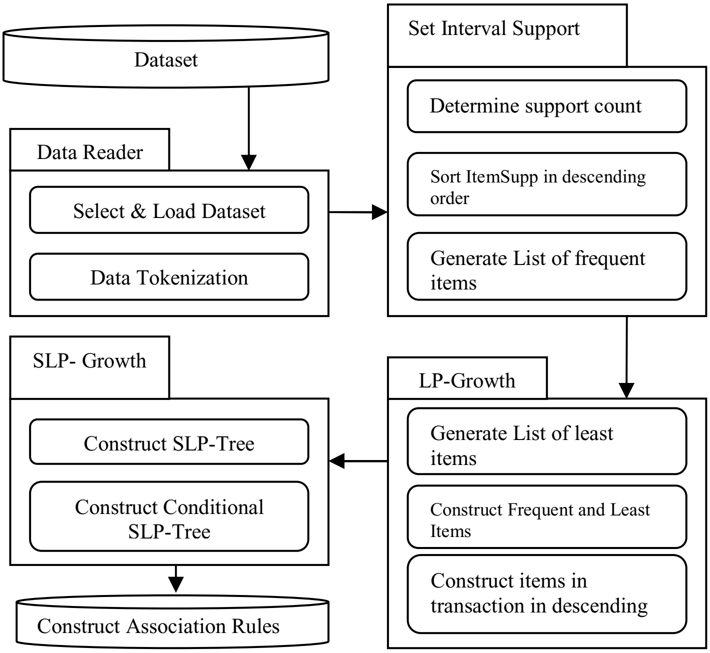

Dataset
Set Interval Support
Determine support count
Data Reader
Sort ItemSupp in descending order
Select & Load Dataset
Generate List of frequent items
Data Tokenization
SLP- Growth
LP-Growth
Construct SLP-Tree
Generate List of least items
Construct Conditional SLP-Tree
Construct Frequent and Least Items
Construct items in transaction in descending
Construct Association Rules

Fig. 2. The Procedure of Mining Critical Least Association Rules

In order to capture the interesting rules and make a decision, the experiment using SLP-Growth method will be conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. The algorithm has been developed using C# as a programming language. The mathematics anxiety dataset used and SLP-Growth produced in this model are in a format of flat file.

6 The Results

We evaluate the proposed algorithm to Mathematics anxiety dataset as in Table 1. To this, we have a dataset comprises the number of transactions (student) is 770 and the number of items (attributes) is 5. Table 3 displays the mapped of original survey dimensions, Likert scale with a new attribute id.

Table 3. The mapping between survey dimensions of mathematic anxiety, Likert scale and a new attribute Id

Survey Dimensions	Likert Scale	Attribute Id
Felt mathematics is difficult	1 – 5	1
Always fail in mathematics	1 – 5	2
Always writing down while mathematic class	1 – 5	3
Anxious if don't understand	1 – 5	4
Lose interest of mathematic	1 – 5	5

Item is constructed based on the combination of survey dimension and its likert scale. For simplicity, let consider a survey dimension "Felt mathematics is difficult" with likert scale "1". Here, an item "11" will be constructed by means of a combination of an attribute id (first characters) and its survey dimension (second character). Different Interval Supports were employed for this experiment.

By embedding FP-Growth algorithm, 2,785 ARs are produced. ARs are formed by applying the relationship of an item or many items to an item (cardinality: many-to-one). Fig. 3 depicts the correlation's classification of interesting ARs. For this dataset, the rule is categorized as significant and interesting if it has positive correlation and CRS value should be equal or greater than 0.5.

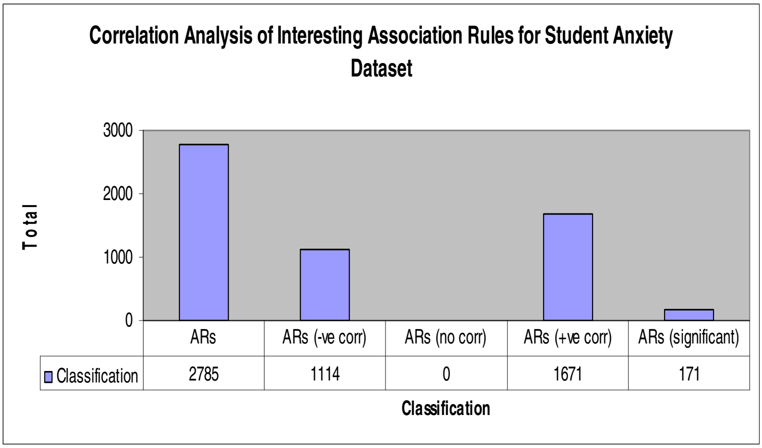


Fig. 3. Classification of ARs using correlation analysis. Only 3.60% from the total of 1,082 ARs are classified as interesting ARs.

Table 4. Top 20 of highest correlation of interesting association rules sorted in descending order of correlation

No.	Association Rules	Supp	Conf	Corr	CRS
1	34 53 23 11 → 41	0.13	100.00	45.29	1.00
2	34 53 11 → 41	0.13	100.00	45.29	1.00
3	34 24 51 15 → 41	0.13	100.00	45.29	1.00
4	34 23 11 → 41	0.13	100.00	45.29	1.00
5	25 51 15 31 → 41	0.13	100.00	45.29	1.00
6	25 15 31 → 41	0.13	100.00	45.29	1.00
7	24 51 15 → 41	0.13	100.00	45.29	1.00
8	55 21 → 31	0.13	100.00	38.50	1.00
9	53 43 25 11 → 31	0.13	100.00	38.50	1.00
10	53 25 11 → 31	0.13	100.00	38.50	1.00
11	53 15 42 → 31	0.13	100.00	38.50	1.00
12	53 12 25 → 31	0.26	100.00	38.50	1.00
13	51 42 21 → 31	0.26	100.00	38.50	1.00
14	51 42 11 21 → 31	0.13	100.00	38.50	1.00
15	45 55 21 → 31	0.13	100.00	38.50	1.00
16	45 15 55 21 → 31	0.13	100.00	38.50	1.00
17	44 53 12 25 → 31	0.26	100.00	38.50	1.00
18	44 23 51 11 → 31	0.13	100.00	38.50	1.00
19	44 23 11 → 31	0.13	100.00	38.50	1.00
20	43 54 22 15 → 31	0.13	100.00	38.50	1.00

Table 4 shows top 20 interesting ARs with numerous types of measurements. The highest correlation value from the selected ARs is 45.29 (No. 1 to 7). From these ARs, there are only one dominant of consequence items, item 41 (Anxious if don't understand is 1 or never). In fact, item 11 only appears 2.21% from the entire dataset. Besides item 41, others consequent item that occur in to 20 interesting ARs is item 31. Item 31 is stand for "Always writing down while mathematic class is 1 or never". For item 31, it occurrences in the dataset is 2.60%. Table 1 also indicates that all interesting ARs have a value of CRS is equal to 1. Therefore, further analysis and study can be used to find out others interesting relationships such as academic performance, personality, attitude, etc. Fig. 4 illustrates the summarization of correlation analysis with different Interval Support.

Fig 5 presents the minimum and maximum values of each employed measure with the confidence value is equal to 100%. The total number of ARs being produced according to the range as stated in Fig 5 is shown in Table 5. The result illustrates that CRS successfully produced the lowest number of ARs as compared to the others measures. The support measure alone is not a suitable measure to be employed to discover the interesting ARs. Although, correlation measure can be used to capture the interesting ARs, it ratio is still 3 times larger than CRS measure. Therefore, CRS is proven to be more efficient and outperformed the benchmarked measures for discovering the interesting ARs.

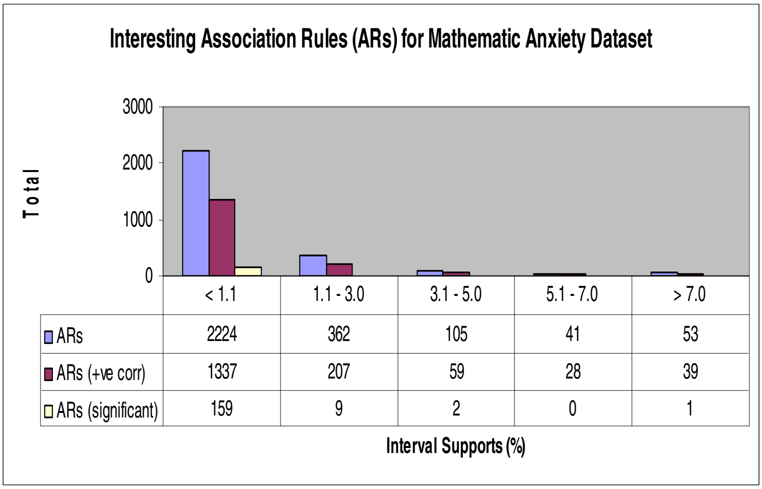


Fig. 4. Correlation analysis of interesting ARs using variety Interval Supports. Generally, total numbers of ARs are decreased when the predefined Interval Supports thresholds are increased.

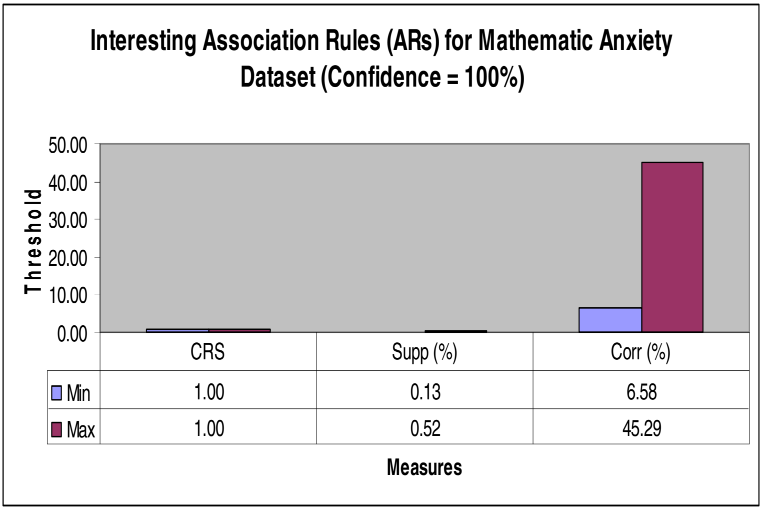


Fig. 5. The range of respective measures based on the confidence value is 100%

Table 5. Summarization of total ARs based on the confidence value is 100%

Measure	Min	Max	Total ARs	Ratio
CRS	1.00	1.00	84	3.80%
Supp	0.13%	0.52%	1,869	84.45%
Corr	6.58%	45.29%	260	11.75%

7 Conclusion

Mathematics anxiety is one of the psychological barriers that students encounter when they are performing a mathematics task [3]. Many mathematics educators find themselves overwhelmed with data, but lack the information they need to make informed decisions. Currently, there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community [4]. One of the popular data mining methods is Association Rules Mining (ARM). In this paper, we had successfully applied an enhanced association rules mining method, so called SLP-Growth (Significant Least Pattern Growth) proposed by [9] for capturing interesting rules in student suffering mathematics anxiety dataset. The dataset was taken from a survey on exploring mathematics anxiety among engineering students in Universiti Malaysia Pahang (UMP). The dataset was taken from a survey on exploring mathematics anxiety among engineering students in Universiti Malaysia Pahang (UMP) [3]. A total 770 students participated in this survey. The respondents were 770 students, consisting of 394 males and 376 females. The respondents are undergraduate students from five engineering faculties at Universiti Malaysia Pahang. It is found that SLP_Growth method is suitable to mine the interesting rules which provide faster and accurate results. Based on the results, educators can obtain recommendation from the rules captured. The results of this research will provide useful information for educators to make a decision on their students more accurately, and to adapt their teaching strategies accordingly. It also can be helpful to assist students in handling their fear of mathematics and useful increasing the quality of learning.

Acknowledgment

The authors would like to thanks Universiti Malaysia Pahang for supporting this work.

References

1. Spielberger, C.D., Vagg, P.R.: Test anxiety: A Transactional Process Model. In: Spielberger, et al. (eds.) Test Anxiety: Theory, Assessment, and Treatment, pp. 1–14. Taylor & Francis, Abington (1995)
2. Anderson, V.: An Online Survey to Assess Student Anxiety and Attitude Response to Six Different Mathematical Problems. In: Watson, J., Beswick, K. (eds.) Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia, pp. 93–102. MERGA Inc. (2007)
3. Vitasari, P., Herawan, T., Wahab, M.N.A., Othman, A., Sinnadurai, S.K.: Exploring Mathematics Anxiety among Engineering' students. Procedia Social and Behavioral Sciences 8, 482–489 (2010)
4. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications 33, 135–146 (2007)
5. Ceglar, A., Roddick, J.F.: Association mining. ACM Computing Surveys 38(2), Article 5, 1–42 (2006)

6. Agrawal, R., Imielinski, T., Swami, A.: Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 5(6), 914–925 (1993)
7. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on the Management of Data, pp. 207–216 (1993)
8. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), pp. 487–499 (1994)
9. Abdullah, Z., Herawan, T., Deris, M.M.: Mining Significant Least Association Rules Using Fast SLP-Growth Algorithm. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN 2010. LNCS, vol. 6059, pp. 324–336. Springer, Heidelberg (2010)
10. Abdullah, Z., Herawan, T., Deris, M.M.: Scalable Model for Mining Critical Least Association Rules. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. LNCS, vol. 6377, pp. 509–516. Springer, Heidelberg (2010)
11. Yun, H., Ha, D., Hwang, B., Ryu, K.H.: Mining Association Rules on Significant Rare Data Using Relative Support. The Journal of Systems and Software 67(3), 181–191 (2003)
12. Zhou, L., Yau, S.: Assocation Rule and Quantative Association Rule Mining Among Infrequent Items. In: The Proceeding of ACM SIGKDD 2007 (2007)
13. Ding, J.: Efficient Association Rule Mining Among Infrequent Items. Ph.D Thesis, University of Illinois at Chicago (2005)
14. Koh, Y.S., Rountree, N.: Finding sporadic rules using apriori-inverse. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 97–106. Springer, Heidelberg (2005)
15. Liu, B., Hsu, W., Ma, Y.: Mining Association Rules With Multiple Minimum Supports. SIGKDD Explorations (1999)
16. Brin, S., Motwani, R., Silverstein, C.: Beyond Market Basket: Generalizing ARs to Correlations. Special Interest Group on Management of Data (SIGMOD), 265-276 (1997)
17. Omniecinski, E.: Alternative Interest Measures For Mining Associations. IEEE Trans. Knowledge and Data Engineering 15, 57–69 (2003)
18. Lee, Y.-K., Kim, W.-Y., Cai, Y.D., Han, J.: CoMine: Efficient Mining Of Correlated Patterns. In: The Proceeding of ICDM 2003 (2003)
19. Han, J., Pei, H., Yin, Y.: Mining Frequent Patterns Without Candidate Generation. In: The Proceeding of SIGMOD 2000. ACM Press, New York (2000)
20. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
21. Au, W.H., Chan, K.C.C.: Mining Fuzzy ARs In A Bank-Account Database. IEEE Transactions on Fuzzy Systems 11(2), 238–248 (2003)
22. Aggrawal, C.C., Yu, P.S.: A New Framework For Item Set Generation. In: Proceedings of the ACMPODS Symposium on Principles of Database Systems, Seattle, Washington (1998)
23. Li, H., Wang, Y., Zhang, D., Zhang, M., Chang, E.Y.: Pfp: Parallel Fp-Growth For Query Recommendation. In: Proceedings of RecSys 2008, pp. 107–114 (2008)
24. Mustafa, M.D., Nabila, N.F., Evans, D.J., Saman, M.Y., Mamat, A.: Association rules on significant rare data using second support. International Journal of Computer Mathematics 83(1), 69–80 (2006)

Methodology for Measuring the Quality of Education Using Fuzzy Logic

Sergio Valdés-Pasarón, Bogart Yail Márquez,
and Juan Manuel Ocegueda-Hernández

Baja California Autonomous University
Calzada Universidad 14418, Tijuana, Baja California, México
{svaldes,bmarquez,jmocegueda}@uabc.edu.mx

Abstract. Fuzzy logic has been considered as a strategy to define the values of the complex realities that define every aspect of education and to demonstrate formalized educational issues. This is because the quality of education has awakened the interest of investigators worldwide because they can be the answer of Education problems, and because some countries spend more resources in funding education compared to others which leads to higher levels of growth. This article proposes a new methodology using fuzzy logic to measure the quality of education by using quantitative and qualitative values with the hopes to develop criteria for the quality of education in a way closer to the realities of Latin American countries.

Keywords: Fuzzy Logic, Education, Economic Growth.

1 Introduction

Human capital theory emerges from the contributions of Mincer (1958) [1], Schultz (1961)[2] and Becker (1964) [3], they considered education as an investment to be made by individuals which allows them the ability to increase their human capital endowment. This investment increases their productivity and, in the neoclassical framework of competitive markets in which this theory is developed, your future income. Thus, establishing a causal relationship between education, productivity and income, so that an increase of education produces a higher level of income and greater economic growth.

In 1960 Gary Becker developed the pure model of human capital, its main hypothesis is based on the fact that a education increases so does the productivity of the individual who receives it [3].

Becker In his study reaches two important conclusions. The first deals with theories of income distribution, rising yields a simple model that emerges from Human capital, this can be described as:

$$Gi=f\ (QNi,\ Ei) \tag{1}$$

Where: G returns, QN innate or natural qualities, E is education or characteristics acquired through investment in human capital, i is a person.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 509–515, 2011.

Becker arrives to an interesting conclusion in this first part of his study and as outlined in his article, "Human Capital." pp.62 and 63, one can assume that there is a whip positive correlation between the natural qualities and the level of educational investment.

In the purpose to satisfy all the expectations and needs of a society as a whole in terms of education this is linked to a series of qualitative and quantitative variables which together gives us an insight as to the differences in quality of Education, by just mentioning various aspects we can refer to: the ratio of students per teacher numbers, the access that students have to technology and by appropriately spending their budget they are able to allocate to the authorities of various countries or the percentage of Gross Domestic Product (GDP) spent on education.

In recent years numerous studies have found that the there is a difficulty in the criteria of variables used to measure elements of education or the factors related to it. These variables include; schooling rates, and the average number of years purchased, for example, the years that an person studies,(what is purchased The number of years a person studies or the average number of years a person has to pay for education but these variables are imperfect measures of the educational component of human capital since they measure the component of quantity, not quality taking a weakening of the value between these comparisons[4-6].

The quality of education has begun to become a high concern among researchers of education because they believe that the expectations and needs of human beings depends on factors like the quality of curricula for which they are prepared, the infrastructure of the country in education, the academic environment which is developed, the faculty and the relationship between teachers and students, among some. Despite this being clearly identified it still remains a difficult task to select the most appropriate indicators to determine which of them have a greater impact on the quality of education [4].

The motivations to incorporate these indicators to improve the quality of education imply that the factors vary from year to year. For instance, in Latin American countries education systems vary widely in terms of the organization of resources, teacher preparation, student-teacher ratio in classrooms, access to technology and education spending per student among other factors, these have high rates of variability among the different countries of Latin America.

The hegemony of the positivist epistemological paradigm in the social sciences has been hindering theoretical constructions that are approximate to reality without reducing their complexity, dismissed with this-not-scientific phenomena such as subjectivity, culture, health, social system and education.

There has recently emerged from different disciplinary fields, a number of theories that come close to the social reality and is able to approach it in all its complexity. These, have a clear epistemological emphasis.

One theory of complexity is that of fuzzy sets as a mathematical formalization of a logical model of imprecise, uncertain, fuzzy, and blurry [7].

2 Measurement of Quality

2.1 How to Measure the Quality of Education

Several factors have been incorporated to measure aspects involving the quality of education, Hanushek and Kim proposed to measure education using skills learned from the test or tests used internationally [8], for example, the Program for International Assessment Students of the OECD (PISA acronym in English) or in the case of Mexico, National Assessment of Academic Achievement in Schools (LINK), which are intended to assess how far students near the end of their compulsory education acquired, to some degree, the knowledge and skills necessary for full participation in society.

The results of these tests show a relationship between the quality of education with the growth of gross national product (GNP) per capita. This suggests that the quality of education is a factor of great importance for the analysis of the relationship between human capital and economic growth[9-11].

However these results have not yet reached a consensus on how to measure qualitative and quantitative jointly due to the heterogeneity in the capture of such data. But, given the difficulty that exists in measuring the quality of education (CE) believes that the main contribution of this work would build a model to measure the quality of education in quantitative and qualitative, eliminating the heterogeneity in the ways of measuring this indicator and reach a final consensus on this controversial issue.

2.2 Fuzzy Logic

The concept of Fuzzy Logic, was conceived by Lotfi Zadeh a professor at the University of California at Berkeley, who disagreed with the classical sets (crisp sets), which allows for only two options; membership or not an item to the all presented as a way of processing information about allowing partial memberships joint as opposed to the classic called Fuzzy Sets [12].

Fuzzy logic versus conventional logic can work with information that is not entirely accurate, so that conventional assessments, which propose that an element always belongs to a certain degree to a set while at the same time never quite belonging to it. This allows for there to be established an efficient way to work with uncertainties, and to put knowledge in the form of rules to a quantitative level, feasible to be processed by computers.

If we make a comparison between the classical and fuzzy logic we can say that classical logic provides a logical parameters between true or false, that is, using binary combinations of 0 and 1, 0 if false and 1 if true. Now if we take into account that the fuzzy logic which introduces a function that expresses the degree of membership of an attribute or variable to a linguistic variable taking the values between 0 and 1, this is called fuzzy set and can be expressed mathematically:

$$A = \{x / \mu A(x) \ \forall \ x \in X\}. \tag{2}$$

Linguistic variable - membership function. By measuring the quality of education, we can analyze the components that make up the whole and if we analyze depending on how much public expenditure there is on education according to classical logic we would be able to know good quality or poor quality regardless of the income distribution of students or the percentage of the generations that come at a higher level, i e if option one then BC = 1 or if the second option MC = 0.

If we use fuzzy logic it is not necessarily the scanned object that has two states because there are other states in which we could label it for example:

Excellent quality VQ = 1, GQ = 0.8 Good quality, medium quality MQ= 0.5, bad quality BQ = 0.1 and too Bad quality TBQ = 0.

3 Methodology

Our methodology is to analyze the indicators used by the United Nations Educational Scientific and Cultural Organization (UNESCO) to measure the education of the countries that have the following input variables to consider.

a) Education expenditure as a percentage of Gross Domestic Product (EXPGDP). This variable represents the percentage of gross domestic product that countries devote to education spending

b) Government Public Expenditure on Education (GPEE). This variable represents the total government spending for education.

c) Distribution of Public Spending by Level of Government (DPSLG). This variable represents the distribution of educational levels (primary, secondary or tertiary) of the total allocation to the education of government expenditure.

d) Pupil- Teacher Ratios (PTR). This variable represents the number of students by teachers at different educational levels.

e) Income rate to last grade of primary (TIUGP). Represents the rate of students who manage to take the primary level in its entirety.

f) Percentage of students who continue to secondary school (SSPE). This variable represents the percent of students who continue their secondary studies completed once a primary school.

g) Expenditure per pupil as a percentage of GDP Per Capita (EPP GDP). Represents the average expenditure per student relative to per capita gross domestic product.
h) Average per pupil expenditure (APPE). Represents the average expenditure per pupil.

3.1 Equations

The relationship between the quality of education and its determinants can be analyzed by a production function of education as:

$$Q=f\,(EF, R) + U \tag{3}$$

Where Q represents the quality of education, EF represents the economic factors ; R resources used in schools and U are unmeasured factors that may affect the quality of education.

The system, which is posed in three blocks is also associated with the process to use fuzzy logic techniques. Blocks are broadcast, and desdifusión inference.

It is important to know the inputs and outputs of the system. The former is formed by the variables and are taken into account in the representation intended by the system. The output is a particular result.

The input variables from a selection process that involves knowing the context of the problem being addressed. As an illustration, the module dealing with MATLAB fuzzy logic, considering the following input and output given the linguistic variables: very bad, bad, average, fair, good, very good, the linguistic variables are made to the perception of education of a particular country.

As input, establishing the factors that influence: EXPGDP, GPEE, DPSLG, PTR, TIUGP, SSPE, EPPGDP and APPE.

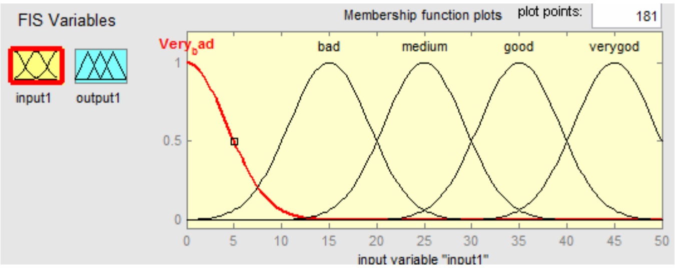


Fig. 1. Membership function

As output, look for the linguistic value that determines a country's education. The functions of each component are:

a) Fuzzification interface; transform variables into fuzzy variables. For this interface must be defined ranges of variation of input variables and fuzzy sets associated with their membership functions.

b) Knowledge Base; contains the linguistic rules of control and information relating to the membership functions of fuzzy sets.

c) Inference engine; makes the task of calculating the output variables from input variables, by the rules of fuzzy inference controller and, delivering output fuzzy sets.

d) Defuzzification interface; gets a diffuse overall output from the aggregation of fuzzy outputs and performs the defuzzification.

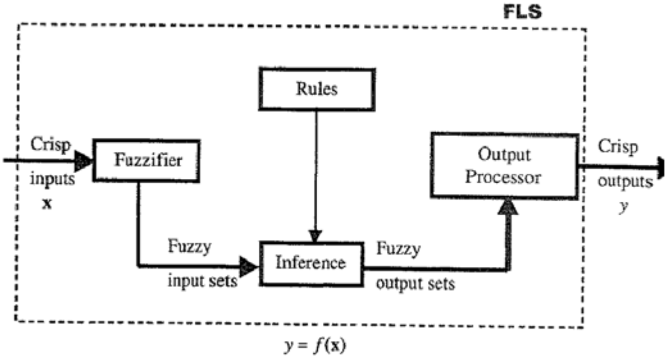


Fig. 2. Fuzzy logic system

To create the fuzzy inference system that calculates the role of education quality, we used the Fuzzy in (Matlab, 2009). Each indicator of education was considered a linguistic variable. Each of these is associated with three fuzzy sets with membership functions of the variable "real" to the set. Each set was labeled with the linguistic labels of "very good", "good", "medium " "bad" and "very bad" to rate the educational value of the indicator considered. The degree of membership of an element in a fuzzy set is determined by a membership function that can take all real values in the range [0.1]. A total of 8 variables defined input and output which corresponds to the quality of education that a country has, it also has the same linguistic labels very good, "" good "," medium "" bad "and" very bad "to describe education.

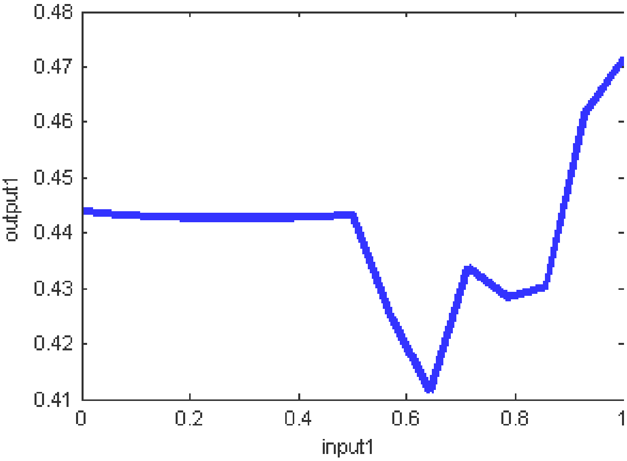


Fig. 3. Resulted ponderous normalized

4 Results

The use of new tools and methodologies for economic theory gives us as output: each level was rated high and low quality of education depending on the limit values "good" and "bad" for all indicators, which do not always coincide with the upper and lower value ranges, it depends if an indicator that is preferred with low or high. The results are shown in Figure 3, where there is the Global Rating education on a value of 1 as the most high. You can see that the higher education is the options if the value of GPEE and EXPGDP is practically the same, while for the other options are appraised lower values being less important technology. In the case of PTR and EPP GDP this show much difference between the cases of low and high quality of education, this is because the ranges of values that can have different indicators are great.

5 Conclusions

The methodology developed for obtaining a function of the quality of education is easy to manage; view and can serve for multiple different sensitivity analysis of changes in the values of education indicators. The great advantage of the methodology based on fuzzy logic is that you can handle an unlimited number of indicators expressed in any unit of measurement. Like any other methodology, it is strongly dependent on the accuracy with which the indicators have been calculated or determined. This paper shows the potential of fuzzy logic to describe particular systems and public policies to determine that the education of a country to be "good "or "bad ".

References

1. Chiswick, B.R., Mincer, J.: Experience and the Distribution of Earnings. Review of Economics of the Household 1(4), 343–361 (2003)
2. Schultz, T.W.: Investment in Human Capital. American Economic Review LI, 1–17 (1961)
3. Becker, G.S.: Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press, Chicago (1964)
4. Lee, J.-W.B., Robert, J.: Schooling Quality in a Cross-Section of Countries. London School of Economics and Political Science (2001)
5. Barro, R.y., Lee, J.-W.: International Comparisons of Educational Attainment. Journal of Monetary Economics 32, 363–394 (1993)
6. Psacharopoulos, G.: Returns to investment in education: a global update. World Development, 22 (1994)
7. Zaded, L.A.: Fuzzy sets. Information and Control 8 (1965)
8. Hanushek, E.A., Kim, D.: Schooling, Labor Force Quality, and Economic Growth. National Bureau of Economic Research (1995)
9. Barro, R.: Human capital and growth in cross-country regressions. Harvard University, Mimeo (1998)
10. Mankiw, G., Romery, D., Weil, D.: A Contribution to the Empirics of Economic Growth. Quartely Journal of Economics, 407–437 (1992)
11. Nelson, R., Phelps, E.: Investment in Humans, Technological Diffusion, and EconomicGrowth. American Economic Review, 69–82 (1966)
12. Zadeh, L.A.: Fuzzy sets and their applications to cognitive and decision processes. In: Zadeh, L.A., et al. (eds.) U.S.-.-J.S.o.F. Sets and U.o.C.B. Their Applications. Academic Press, New York (1975)

Rough Set Theory for Feature Ranking of Traditional Malay Musical Instruments Sounds Dataset

Norhalina Senan[1], Rosziati Ibrahim[1], Nazri Mohd Nawi[1],
Iwan Tri Riyadi Yanto[2], and Tutut Herawan[3]

[1] Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia,
Batu Pahat 86400, Johor, Malaysia
{halina,rosziati,nazri}@uthm.edu.my
[2] Department of Mathematics
Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia
iwan015@gmail.com
[3] Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang, Lebuh Raya Tun Razak,
Gambang 26300, Pahang, Malaysia
tutut@ump.edu.my

Abstract. This paper presents an alternative feature ranking technique for Traditional Malay musical instruments sounds dataset using rough-set theory based on the maximum degree of dependency of attributes. The modeling process comprises seven phases: data acquisition, sound editing, data representation, feature extraction, data discretization, data cleansing, and finally feature ranking using the proposed technique. The results show that the selected features generated from the proposed technique able to reduce the complexity process.

Keywords: Rough Set Theory; Dependency of attribute; Feature ranking, Traditional Malay musical instruments sounds dataset.

1 Introduction

With the growing volume of digital audio data and feature schemes, feature ranking has become very vital aspect in musical instruments sounds classification problems. In general, the purpose of the feature ranking is to alleviate the effect of the '*curse of dimensionality*'. While, from the classification point of view, the main idea of feature ranking is to construct an efficient and robust classifier. It has been proven in practice that the optimal classifier difficult to classify accurately if the poor features are presented as the input. This is because some of the input features have poor capability to split among different classes and some are highly correlated [1]. As a consequence, the overall classification performance might decrease with this large number of features available. For that, finding only relevant subset of features may significantly reduce the complexity process and improve the classification performance by eliminating irrelevant and redundant features.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 516–529, 2011.

This shows that the problem of feature ranking must be addressed appropriately. For that, various feature ranking algorithms in musical instruments sounds classification have been proposed by several researchers [1,2,3]. Liu and Wan [1] carried out a study on classifying the musical instruments into five families (brass, keyboard, percussion, string and woodwind) using NN, k-NN and Gaussian mixture model (GMM). Three categories of features schemes which are temporal features, spectral features and coefficient features (with total of 58 features) were exploited. A sequential forward selection (SFS) is used to choose the best features. The k-NN classifier using 19 features achieves the highest accuracy of 93%. In Deng et. al [2], they conducted a study on selecting the best features schemes based on their classification performance. The 44 features from three categories of features schemes which are human perception, cepstral features and MPEG-7 were used. To select the best features, three entropy-based feature selection techniques which are Information Gain, Gain Ratio and Symmetrical Uncertainty were utilized. The performance of the selected features was assessed and compared using five classifiers which are k-nearest neighbor (*k*-NN), naive bayes, support vector machine (SVM), multilayer perceptron (MLP) and radial basic functions (RBF). Deng et al [[2] demonstrated that the Information Gain produce the best classification accuracy up to 95.5% for the 20 best features with SVM and RBF classifiers. Benetos *et. al* [3] applied subset selection algorithm with branch-bound search strategy for feature reduction. A combination of 41 features from general audio data, MFCC and MPEG-7 was used. By using the best 6 features, the non-negative matrix factorization (NMF) classifier yielded an accuracy rate of 95.2% at best. They found that the feature subset selection method adopted in their study able to increase the classification accuracy. In overall, all these works demonstrate that the reduced features able to produce highest classification rate with less computational time. On the other hand, Deng *et al.* [2] claimed that benchmarking is still an open issue in this area of research. This shows that the existing feature ranking approaches applied in the various sound files may not effectively work to other conditions. Therefore, there were significant needs to explore other feature ranking methods with different types of musical instruments sounds in order to find the best solution.

One of the potential techniques for dealing with this problem is based on the rough set theory. The theory of rough set proposed by Pawlak in 1980s [4] is a mathematical tool for dealing with the vague and uncertain data. Rough sets theory is one of the useful tools for feature selection [5,6,7]. Banerjee *et al.* [5] claimed that the concept of *core* in rough set is relevant in feature selection to identify the essential features amongst the non-redundant ones. The attractive characteristics of rough set in tackling the problem of imprecision, uncertainty, incomplete, irrelevant or redundancy in the large dataset, has magnificently attracted researchers in wide areas of data mining domain to utilize rough set for feature selection. However, to date, a study on rough sets for feature ranking of musical instruments sounds classification is scarce and still needs an intensive research. It is well-known that one of the most crucial aspects of musical instruments sounds classification is to find the best features schemes. With the special capability of rough set for feature ranking, we are going to apply this technique in musical instruments sounds classification to overcome this issue.

In this paper, an alternative feature selection technique based on rough set theory for Traditional Malay musical instruments sounds classification is proposed. This technique is developed based on rough set approximation using Maximum Degree of dependency of Attributes (MDA) technique proposed by [8]. The idea of this technique is to choose the most significant features by ranking the relevant features based on the highest dependency of attributes on the dataset and then remove the redundant features with the similar dependency value. To accomplish this study, the quality of the instruments sounds is first examined. Then, the 37 features from two combination of features schemes which are perception-based and Mel-Frequency Cepstral Coefficients (MFCC) are extracted [9]. In order to employ the rough set theory, this original dataset (continuous values) is then discritized into categorical values by using equal width and equal frequency binning algorithm [10]. Afterwards, data cleansing process is done to remove the irrelevant features. Finally, the proposed technique is then adopted to rank and select the best feature set from the large number of features available in the dataset.

The rest of this paper is organized as follows: Section 2 presents the theory of rough set. Section 3 describes the details of the modelling process. A discussion of the result is presented in Section 4 followed by the conclusion in Section 5.

2 Rough Set Theory

In this Section, the basic concepts of rough set theory in terms of data are presented.

2.1 Information System

Data are often presented as a table, columns of which are labeled by *attributes*, rows by *objects* of interest and entries of the table are *attribute values*. By an *information system*, a 4-tuple (quadruple) $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute *a*, $f : U \times A \to V$ is a total function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, called information (knowledge) function. In many applications, there is an outcome of classification that is known. This *a posteriori* knowledge is expressed by one (or more) distinguished attribute called decision attribute; the process is known as supervised learning. An information system of this kind is called a decision system. A *decision system* is an information system of the form $D = (U, A \bigcup \{d\}, V, f)$, where $d \notin A$ is the *decision attribute*. The elements of *A* are called *condition attributes*.

2.2 Indiscernibility Relation

The notion of indiscernibility relation between two objects can be defined precisely.

Definition 2.1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ are said to be B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x,a) = f(y,a)$, for every $a \in B$.*

Obviously, every subset of *A* induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute *B*, denoted by $IND(B)$, is an equivalence relation. It is well known that, an equivalence relation induces unique partition. The partition of *U* induced by $IND(B)$ in $S = (U, A, V, f)$ denoted by U / B and the equivalence class in the partition U / B containing $x \in U$, denoted by $[x]_B$.

Given arbitrary subset $X \subseteq U$, in general, *X* as union of some equivalence classes in *U* might be not presented. It means that, it may not be possible to describe *X* precisely in AS . X might be characterized by a pair of its approximations, called lower and upper approximations. It is here that the notion of rough set emerges.

2.3 Set Approximations

The indiscernibility relation will be used next to define approximations, basic concepts of rough set theory. The notions of lower and upper approximations of a set can be defined as follows.

Definition 2.2. *Let $S = (U, A, V, f)$ be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by $\underline{B}(X)$ and B-upper approximations of X, denoted by $\overline{B}(X)$, respectively, are defined by*

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\}.$$

The accuracy of approximation (accuracy of roughness) of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|},$$

where $|X|$ denotes the cardinality of *X*. For empty set ϕ, $\alpha_B(\phi) = 1$ is defined. Obviously, $0 \leq \alpha_B(X) \leq 1$. If *X* is a union of some equivalence classes of *U*, then $\alpha_B(X) = 1$. Thus, the set *X* is *crisp* (precise) with respect to *B*. And, if *X* is not a union of some equivalence classes of *U*, then $\alpha_B(X) < 1$. Thus, the set *X* is *rough* (imprecise) with respect to *B* [11]. This means that the higher of accuracy of approximation of any subset $X \subseteq U$ is the more precise (the less imprecise) of itself.

Another important issue in database analysis is discovering dependencies between attributes. Intuitively, a set of attributes *D* depends totally on a set of attributes *C*, denoted $C \Rightarrow D$, if all values of attributes from *D* are uniquely determined by values of attributes from *C*. In other words, *D* depends totally on *C*, if there a functional dependency between values of *D* and *C*. The formal definition of attributes dependency is given as follows.

Definition 2.3. *Let* $S = (U, A, V, f)$ *be an information system and let D and C be any subsets of A. Attribute D is functionally depends on C , denoted* $C \Rightarrow D$*, if each value of D is associated exactly one value of C.*

2.4 Dependency of Attributes

Since information system is a generalization of a relational database. A generalization concept of dependency of attributes, called a *partial dependency* of attributes is also needed.

Definition 2.4. *Let* $S = (U, A, V, f)$ *be an information system and let D and C be any subsets of A. The dependency attribute D on C in a degree k* $(0 \le k \le 1)$*, is denoted by* $C \Rightarrow_k D$*, where*

$$k = \gamma(C, D) = \frac{\sum_{X \in U/D} \left|\underline{C}(X)\right|}{|U|}. \tag{1}$$

Obviously, $0 \le k \le 1$. If all set X are crisp, then $k = 1$. The expression $\sum_{X \in U/D} \left|\underline{C}(X)\right|$, called a lower approximation of the partition U / D with respect to *C*, is the set of all elements of *U* that can be uniquely classified to blocks of the partition U / D, by means of *C*. *D* is said to be fully depends (in a degree of *k*) on *C* if $k = 1$. Otherwise, *D* is partially depends on *C*. Thus, *D* fully (partially) depends on *C*, if all (some) elements of the universe *U* can be uniquely classified to equivalence classes of the partition U / D, employing *C*.

2.3 Reducts and Core

A *reduct* is a minimal set of attributes that preserve the indiscernibility relation. A *core* is the common parts of all reducts. In order to express the above idea more precisely, some preliminaries definitions are needed.

Definition 2.5. *Let* $S = (U, A, V, f)$ *be an information system and let B be any subsets of A and let a belongs to B. It say that a is dispensable (superfluous) in B if* $U / (B - \{b\}) = U / B$*, otherwise a is indispensable in B.*

To further simplification of an information system, some dispendable attributes from the system can be eliminated in such a way that the objects in the table are still able to be discerned as the original one.

Definition 2.6. *Let* $S=(U,A,V,f)$ *be an information system and let B be any subsets of A. B is called independent (orthogonal) set if all its attributes are indispensable.*

Definition 2.7. *Let* $S=(U,A,V,f)$ *be an information system and let B be any subsets of A. A subset* B^* *of B is a reduct of B if* B^* *is independent and* $U/B^*=U/B$.

Thus a reduct is a set of attributes that preserves partition. It means that a reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe. While computing equivalence classes is straighforward, but the problem of finding minimal reducts in information systems is NP-hard. Reducts have several important properties. One of them is a *core*.

Definition 2.8. *Let* $S=(U,A,V,f)$ *be an information system and let B be any subsets of A. The intersection off all reducts of is called the core of B, i.e.,*

$$\text{Core}(B)=\bigcap\text{Red}(B),$$

Thus, the *core* of *B* is the set off all indispensable attributes of *B*. Because the core is the intersection of all reducts, it is included in every reduct, i.e., each element of the core belongs to some reduct. Thus, in a sense, the core is the most important subset of attributes, for none of its elements can be removed without affecting the classification power of attributes.

3 The Modeling Process

In this section, the modelling process of this study is presented. There are seven main phases which are data acquisition, sound editing, data representation, feature extraction, data discretization, data cleansing and feature ranking using proposed technique known as 'Feature Selection using Dependency Attribute' (FSDA). Figure 1 illustrates the phases of this process. To conduct this study, the proposed model is implemented in MATLAB version 7.6.0.324 (R2008a). It is executed on a processor Intel Core 2 Duo CPUs. The total main memory is 2 gigabytes and the operating system is Windows Vista. The details of the modelling process as follows:

3.1 Data Acquisition, Sound Editing, Data Representation and Feature Extraction

The 150 sounds samples of Traditional Malay musical instruments were downloaded from personal [15] and *Warisan Budaya Malaysia* web page [16]. The dataset comprises four different families which are membranophones, idiophones, aerophones and chordophones. This original dataset is non-benchmarking (real work) data. The number of the original sounds per family is imbalance which also differs in term of the lengthwise. It is well-known that the quality of the data is one of the factors that might affect the overall classification task. To this, the dataset is firstly edited and trimmed. Afterwards, two categories of features schemes which are perception-based and MFCC features were extracted. All 37 extracted features from these two categories are shown in Table 1. The first 1-11 features represent the perception-based features and 12-37 are MFCC's features. The mean and standard deviation were then calculated for each of these features. In order to avoid biased classification, the dataset are then eliminated to uniform size. The details of these phases can be found in [17].

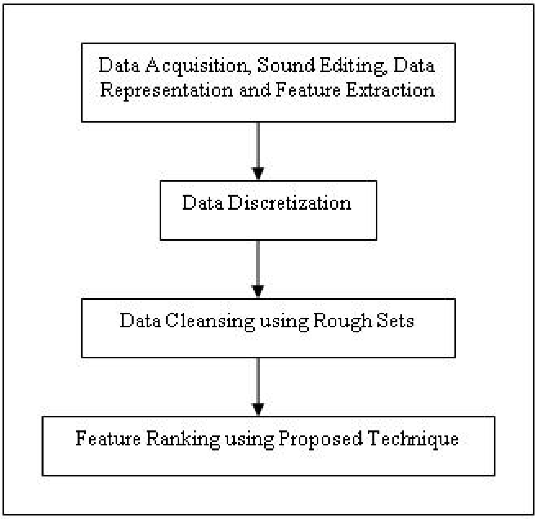

Data Acquisition, Sound Editing, Data Representation and Feature Extraction
Data Discretization
Data Cleansing using Rough Sets
Feature Ranking using Proposed Technique

Fig. 1. The modelling process for feature ranking of the Traditional Malay musical instruments sounds classification

Table 1. Features Descriptions

Number	Description
1	Zero Crossing
2-3	Mean and Standard Deviation of Zero Crossings Rate
4-5	Mean and Standard Deviation of Root-Mean-Square
6-7	Mean and Standard Deviation of Spectral Centroid
8-9	Mean and Standard Deviation of Bandwidth
10-11	Mean and Standard Deviation of Flux
12-37	Mean and Standard Deviation of the First 13 MFCCs

3.2 Data Discretization

The features (attributes) extracted in the dataset are in the form of continuous value with non-categorical features (attributes). In order to employ the rough set approach in the proposed technique, it is essential to transform the dataset into categorical ones. For that, the discretization technique known as the equal width binning in [10] is applied. In this study, this unsupervised method is modified to be suited in the classification problem. The algorithm first sort the continuous valued attribute, then the minimum $x_{\min}$ and the maximum $x_{\max}$ of that attribute is determined. The interval width, *w*, is then calculated by:

$$w = \frac{x_{\max} - x_{\min}}{k^{*}},$$

where, $k*$ is a user-specified parameter for the number of intervals to discretize of each target class. The interval boundaries are specified as $x_{\min} + w_i$, where $i = 1,2,\cdots,k-1$. Afterwards, the equal frequency binning method is used to divide the sorted continuous values into *k* interval where each interval contains approximately n/k data instances with adjacent values of each class. In this study, the difference of *k* value (from 2 to 10) is examined. The purpose is to identify the best *k* value which able to produce highest classification rate. For that, rough set classifier is used.

3.3 Data Cleansing Using Rough Set

As mentioned in Section 1, the dataset used in this study are raw data obtained from multiple resources (non-benchmarking data). In sound editing and data representation phases, the reliability of the dataset used have been assessed. However, the dataset may contain irrelevant features. Generally, the irrelevant features present in the dataset are features that having no impact on processing performance. However, the existence of these features in the dataset might increase the response time. For that, in this phase, the data cleansing process based on rough sets approach explained in subsection 2.5 is performed to eliminate the irrelevant features from the dataset.

3.4 The Proposed Technique

In this phase, the construction of the feature ranking technique using rough set approximation in an information system based on dependency of attributes is presented. The idea of this technique is derived from [8]. The relation between the properties of roughness of a subset $X \subseteq U$ with the dependency between two attributes is firstly presented as in Proposition 3.1.

Proposition 3.1. *Let $S = (U, A, V, f)$ be an information system and let D and C be any subsets of A. If D depends totally on C, then*

$$\alpha_D(X) \le \alpha_C(X),$$

for every $X \subseteq U$.

Proof. Let *D* and *C* be any subsets of *A* in information system $S=(U,A,V,f)$. From the hypothesis, the inclusion $IND(C)\subseteq IND(D)$ holds. Furthermore, the partition U/C is finer than that U/D, thus, it is clear that any equivalence class induced by $IND(D)$ is a union of some equivalence class induced by $IND(C)$. Therefore, for every $x\in X\subseteq U$, the property of equivalence classes is given by

$$[x]_C\subseteq[x]_D\,.$$

Hence, for every $X\subseteq U$, we have the following relation

$$\underline{D}(X)\subseteq\underline{C}(X)\subset X\subset\overline{C}(X)\subseteq\overline{D}(X).$$

Consequently,

$$\alpha_D(X)=\frac{|\underline{D}(X)|}{|\overline{D}(X)|}\leq\frac{|\underline{C}(X)|}{|\overline{C}(X)|}=\alpha_C(X).$$

The generalization of Proposition 3.1 is given below.

Proposition 3.2. *Let $S=(U,A,V,f)$ be an information system and let $C_1,C_2,\cdots,C_n$ and D be any subsets of A. If $C_1\Rightarrow_{k_1}D, C_2\Rightarrow_{k_2}D,\cdots,C_n\Rightarrow_{k_n}D$, where $k_n\leq k_{n-1}\leq\cdots\leq k_2\leq k_1$, then*

$$\alpha_D(X)\leq\alpha_{C_n}(X)\leq\alpha_{C_{n-1}}(X)\leq\cdots\leq\alpha_{C_2}(X)\leq\alpha_{C_1}(X),$$

for every $X\subseteq U$.

Proof. Let $C_1,C_2,\cdots,C_n$ and *D* be any subsets of *A* in information system *S*. From the hypothesis and Proposition 3.1, the accuracies of roughness are given as

$$\begin{gathered}\alpha_D(X)\leq\alpha_{C_1}(X)\\ \alpha_D(X)\leq\alpha_{C_2}(X)\\ \vdots\\ \alpha_D(X)\leq\alpha_{C_n}(X)\end{gathered}$$

Since $k_n\leq k_{n-1}\leq\cdots\leq k_2\leq k_1$, then

$$\begin{gathered}[x]_{C_n}\subseteq[x]_{C_{n-1}}\\ [x]_{C_{n-1}}\subseteq[x]_{C_{n-2}}\\ \vdots\\ [x]_{C_2}\subseteq[x]_{C_1}\,.\end{gathered}$$

Obviously,

$$\alpha_D(X)\le\alpha_{C_n}(X)\le\alpha_{C_{n-1}}(X)\le\cdots\le\alpha_{C_2}(X)\le\alpha_{C_1}(X). \qquad \square$$

Figure 2 shows the pseudo-code of the proposed technique. The technique uses the dependency of attributes in the rough set theory in information systems. It consists of five main steps. The first step deals with the computation of the equivalence classes of each attribute (feature). The equivalence classes of the set of objects U can be obtained using the indiscernibility relation of attribute $a_i \in A$ in information system $S=(U,A,V,f)$. The second step deals with the determination of the dependency degree of attributes. The degree of dependency attributes can be determined using formula in Equation (1). The third step deals with selecting the maximum dependency degree. Next step, the attribute is ranked with the ascending sequence based on the maximum of dependency degree of each attribute. Finally, all the redundant attributes are identified. The attribute with the highest value of the maximum degree of dependency within these redundant attributes is then selected.

```
Algorithm: FSDA
Input: Data set with categorical value
Output: Selected non-redundant attribute
Begin
    Step 1. Compute the equivalence classes using the
    indiscernibility relation on each attribute.
    Step 2. Determine the dependency degree of attribute $a_i$
    with respect to all $a_j$, where $i \neq j$.
    Step 3. Select the maximum of dependency degree of each
    attribute.
    Step 4. Rank the attribute with ascending sequence based
    on the maximum of dependency degree of each attribute.
    Step 5. Select the attribute with the highest value of
    maximum degree of dependency within the redundant
    attributes.
End
```

Fig. 2. The FSDA algorithm

4 Results and Discussion

The main objective of this study is to select the best features using the proposed technique. Afterwards, the performance of the selected features is assessed using two different classifiers which are rough set and MLP. As mentioned, the assessment of the performance is based on the accuracy rate and response time achieved. Thus, in this section, the results of this study are presented as follows:

4.1 The Best *k* Value for Discretization Is Determined

The original dataset in continuous value is discritized into categorical form in order to employ rough set theory. For that, the modified equal width binning technique is employed. In this study, the difference of *k* (number of intervals) value from 2 to 10 is also investigated. The best *k* value is determined based on the highest classification accuracy achieved by the rough set classifier. The finding reveals that *k*=3 able to generate the highest classification accuracy up to 99% as shown in Table 2. This *k* value is then applied in the propose feature ranking technique to identify the best features for the dataset.

4.2 Irrelevant Features Is Eliminated

The dataset is represented in decision table form as $S = (U, A \cup \{d\}, V, f)$. There are 1116 instances in the universe *U*, with the family of the instruments as the decision attribute *d* and all other attributes shown in Table 1 as the set of condition attributes, *A*. The distribution of all instances in each class is uniform with no missing values in the data. From the data cleansing step, it is found that {MMFCC1, SMFCC1} is the dispensable (irrelevant) set of features. It is means that the number of the relevant features is 35 out of 37 of original full features. Thus, this relevant features can be represented as A−{MMFCC1, SMFCC1}.

4.3 Finding the Best Features

In this experiment, the proposed technique is employed to identify the best features for Traditional Malay musical instruments sounds classification. As demonstrated in Table 3, all the 35 relevant features are ranked in ascending sequence based on the value of the maximum degree of attribute dependency. From the table, it is fascinating to see that some of the features adopted in this study are redundant. In order to reduce the dimensionality of the dataset, only one of these redundant features is selected. It is revealed that the proposed feature ranking technique able to select the best 17 features out of 35 features available successfully. The best selected features are given in Table 4.

Table 2. Finding the best *k* value for discretization

k	2	3	4	5	6	7	8	9	10
Classification Accuracy (%)	93.6	98.9	98.6	98.4	98.4	98.3	98.3	98.3	98.3

Table 3. Feature ranking using proposed method

Number of Features	Name of Features	Maximum Degree of Dependency of Attributes
3	STDZCR	0.826165
36	SMFCC12	0.655914
23	MMFCC12	0.52509
24	MMFCC13	0.52509
22	MMFCC11	0.237455
30	SMFCC6	0.208781
31	SMFCC7	0.208781
1	ZC	0.193548
37	SMFCC13	0.1819
32	SMFCC8	0.108423
33	SMFCC9	0.108423
34	SMFCC10	0.108423
35	SMFCC11	0.108423
27	SMFCC3	0.087814
29	SMFCC5	0.087814
11	STDFLUX	0.077061
21	MMFCC10	0.077061
20	MMFCC9	0.074373
6	MEANC	0.065412
19	MMFCC8	0.065412
18	MMFCC7	0.056452
28	SMFCC4	0.056452
7	STDC	0.042115
8	MEANB	0.042115
9	STDB	0.042115
13	MMFCC2	0.031362
16	MMFCC5	0.031362
17	MMFCC6	0.031362
5	STDRMS	0.021505
10	MEANFLUX	0.011649
2	MEANZCR	0
4	MEANRMS	0
14	MMFCC3	0
15	MMFCC4	0
26	SMFCC2	0

Table 4. The selected features

Number of Features	Name of Features	Maximum Degree of Dependency of Attributes
3	STDZCR	0.826165
36	SMFCC12	0.655914
23	MMFCC12	0.52509
22	MMFCC11	0.237455
30	SMFCC6	0.208781
1	ZC	0.193548
37	SMFCC13	0.1819
32	SMFCC8	0.108423
27	SMFCC3	0.087814
11	STDFLUX	0.077061
20	MMFCC9	0.074373
6	MEANC	0.065412
18	MMFCC7	0.056452
7	STDC	0.042115
13	MMFCC2	0.031362
5	STDRMS	0.021505
10	MEANFLUX	0.011649

5 Conclusion

In this study, an alternative technique for feature ranking using rough set theory based on the maximum dependency of the attributes Traditional Malay musical instruments sounds is proposed. A non-benchmarking dataset of Traditional Malay musical instruments sounds is utilized. Two categories of features schemes which are perception-based and MFCC that consist of 37 attributes are extracted. Afterward, the dataset is discretized into 3 categorical values. Finally, the proposed technique is then adopted for feature ranking through feature ranking and dimensionality reduction.

In overall, the finding shows that the relevant features selected from the proposed model able to reduce the complexity. Thus, the future work will investigate the process of classifier techniques to evaluate the performance of the selected features in terms of the accuracy rate and response time produced.

Acknowledgement

This work was supported by Universiti Tun Hussein Onn Malaysia (UTHM).

References

1. Liu, M., Wan, C.: Feature Selection for Automatic Classification of Musical Instrument Sounds. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2001, pp. 247–248 (2001)
2. Deng, J.D., Simmermacher, C., Cranefield, S.: A Study on Feature Analysis for Musical Instrument Classification. IEEE Transactions on System, Man, and Cybernetics-Part B: Cybernetics 38(2), 429–438 (2008)
3. Benetos, E., Kotti, M., Kotropoulus, C.: Musical Instrument Classification using Non-Negative Matrix Factorization Algorithms and Subset Feature Selection. In: Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, vol. 5, pp. 221–224 (2006)
4. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Science 11, 341–356 (1982)
5. Banerjee, M., Mitra, S., Anand, A.: Feature Selection using Rough Sets. In: Banerjee, M., et al. (eds.) Multi-Objective Machine Learning. SCI, vol. 16, pp. 3–20 (2006)
6. Modrzejewski, M.: Feature Selection using Rough Sets Theory. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 213–226. Springer, Heidelberg (1993)
7. Li, H., Zhang, W., Xu, P., Wang, H.: Rough Set Attribute Reduction in Decision Systems. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2008. LNCS (LNAI), vol. 5097, pp. 132–141. Springer, Heidelberg (2008)
8. Herawan, T., Mustafa, M.D., Abawajy, J.H.: Rough set approach for selecting clustering attribute. Knowledge Based Systems 23(3), 220–231 (2010)
9. Senan, N., Ibrahim, R., Nawi, N.M., Mokji, M.M.: Feature Extraction for Traditional Malay Musical Instruments Classification. In: Proceeding of International Conference of Soft Computing and Pattern Recognition, SOCPAR 2009, pp. 454–459 (2009)
10. Palaniappan, S., Hong, T.K.: Discretization of Continuous Valued Dimensions in OLAP Data Cubes. International Journal of Computer Science and Network Security 8, 116–126 (2008)
11. Pawlak, Z.: Rough set and Fuzzy sets. Fuzzy sets and systems 17, 99–102 (1985)
12. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Science 177(1), 3–27 (2007)
13. Zhao, Y., Luo, F., Wong, S.K.M., Yao, Y.: A general definition of an attribute reduct. In: Yao, J., Lingras, P., Wu, W.-Z., Szczuka, M.S., Cercone, N.J., Ślęzak, D. (eds.) RSKT 2007. LNCS (LNAI), vol. 4481, pp. 101–108. Springer, Heidelberg (2007)
14. Pawlak, Z.: Rough classification. International Journal of Human Computer Studies 51, 369–383 (1983)
15. Warisan Budaya Malaysia: Alat Muzik Tradisional, `http://malaysiana.pnm.my/kesenian/Index.htm`
16. Shriver, R.: Webpage, `http://www.rickshriver.net/hires.htm`
17. Senan, N., Ibrahim, R., Nawi, N.M., Mokji, M.M., Herawan, T.: The Ideal Data Representation for Feature Extraction of Traditional Malay Musical Instrument Sounds Classification. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2010. LNCS, vol. 6215, pp. 345–353. Springer, Heidelberg (2010)

Pi-Sigma Neural Network for Temperature Forecasting in Batu Pahat

Noor Aida Husaini[1], Rozaida Ghazali[1],
Nazri Mohd Nawi[1], and Lokman Hakim Ismail[2]

[1] Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johore, Malaysia
[2] Faculty of Environmental and Civil Engineering,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johore, Malaysia
gi090003@siswa.uthm.edu.my,
{rozaida,nazri,lokman}@uthm.edu.my

Abstract. In this study, two artificial neural network (ANN) models, a Pi-Sigma Neural Network (PSNN) and a three-layer multilayer perceptron (MLP), are applied for temperature forecasting. PSNN is use to overcome the limitation of widely used MLP, which can easily get stuck into local minima and prone to overfitting. Therefore, good generalisation may not be obtained. The models were trained with backpropagation algorithm on historical temperature data of Batu Pahat region. Through 810 experiments, we found that PSNN performs considerably better results compared to MLP for daily temperature forecasting and can be suitably adapted to forecasts a particular region using the historical data over larger geographical areas.

Keywords: pi-sigma neural network, temperature forecasting, backpropagation.

1 Introduction

Malaysia's weather is generally hot and sunny all year around [1] where an average daily temperature is about ±32ºC (±90ºF) during day time and falls to merely ±24ºC (±75ºF) at night. The temperature is affected by humidity, which is consistent in the range of 75% to 80% throughout the year [2]. While the average annual rainfall of around 200 cm to 300 cm, thus the days are typically warm whilst the nights and the early mornings are moderately cool [1]. Temperature which can be considered as a kind of atmospheric time series data, involve the time index on a predetermined or countably unlimited set of values. Indeed, temperature can be defined qualitatively as a measure of hotness which can be categorised in a sequence according to their hotness [3]. Since it is a stochastic process, these values usually comprise measurements of a physical system took on a specific time delays that might be hours, days, months or years. Accurate measurement of the temperature is highly difficult to fulfill. Some great observations are needed to obtain accuracies for the temperature measurement [4].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 530–541, 2011.

Temperature forecasting is one of the most important components in successful operation in any weather forecasting system. The greatest interest in developing methods for more accurate predictions for temperature forecasting has led to the development of several methods that can be mull over into three categories: 1) physical methods that utilise the laws of radiation, 2) statistical-empirical methods that consist of mathematical approaches fitted to archive meteorological data and 3) numerical-statistical methods [5], [6] with the arrival of numerical weather modelling. This method has been adopted by many Meteorological Services [5], which later on is use as input to obtain a temperature forecast. There have been many different scientific efforts in order to realise better results in the domain of forecasting meteorological parameters [5], [6], [7] like temperature. Temperature forecasting which usually done using the projected images of data taken by meteorological satellites to assess future trends [7], is intrinsically costlier and only proficient of providing certain information. Nevertheless, the extensive use of such numerical weather method is still restricted by the availability of numerical weather prediction products, thus leading to various studies being conducted for temperature forecasting. Moreover, most meteorological processes often exhibit temporal scale in parameter estimation and spatial variability [8].

Therefore, a variety of artificial neural network (ANN) configurations has been developed. The predictive potentiality of ANN is widely acknowledged and applicable to many range of problems, including simulation and forecasting of meteorological parameters. It can be said that ANN provides an attractive alternative tool and has the ability to approximate the uncertainty that relates the forecast data to the actual data [8]. Furthermore, it takes into account for non-linearity of meteorological processes that are difficult to solve by the aforementioned techniques. Given the reported successes of ANN applications in forecasting, it would appear that a suitably designed ANN might be able to represents the flow of temperature. The Multilayer Perceptron (MLP), which is a counterpart of conventional ANN, has been increasingly popular among researchers as various approaches have been applied in temperature forecasting. This includes the work of Hayati and Mohebi [9] for short-term temperature forecasting in Kermanshah, Iran. The results showed that MLP performed much better when it achieved the minimum forecasting error and reasonable forecasting accuracy. Lee *et al.* [10] proposed a new method for temperature prediction to improve the rate of forecasting accuracy using high-order fuzzy logical relationships. The method was implemented by adjusting the length of each interval in the universe of discourse. Smith *et al.* [11] presents an application of ANN for air temperature predictions based on near real-time data with the reduction of prediction error. The prediction error is achieved by increasing the number of distinct observations in the training set.

A similar problem has also been tackled by Radhika *et al.* [12] using Support Vector Machines (SVM) for one-step-ahead weather prediction applications. It is found that SVM consistently gives better results compared to MLP when daily maximum temperature for a span of previous n days was used as the input of the network [12]. Wang and Chen [13] provided the prediction of daily temperature for Taiwan using automatic clustering techniques. The techniques used to cluster the historical numerical data into intervals of different lengths based on two-factor high-order fuzzy time series. Baboo and Shereef [14] forecast temperature using real time dataset and compare it with practical working of meteorological department. Results showed that the convergence analysis is improved by using simplified conjugate gradient (CG) method.

Though, MLP is the most common type of network in use and its ability in forecasting has been well performed, it is however, MLP is unable to handle non-smooth, discontinuous training data, and complex mappings [15], [16]. Somehow, MLP that can be thought as a blackbox, whereas taking in and giving out information [7] is unable to explain the output of meteorological processes obviously. Indeed, MLP is prone to overfit the data [12]. On the other hand, MLP also suffer long training times and often reach local minima [15].

This is then; be the first motivation of using Pi-Sigma Neural Network (PSNN) [17] in this study. PSNN is an openbox model whereas each neuron and weights are mapped to function variable and coefficient. The second motivation of using PSNN is due to the network model that can automatically select the initial coefficients for nonlinear data analysis. Despite, this network model possess high learning capabilities [15] that reduce the complexity of the network's structure in terms of less weights and nodes required [17]. For that reason, PSNN was chosen to learn the historical temperature data in the study area, namely Batu Pahat for daily temperature forecasting, and is benchmarked with MLP.

This paper is structured as follows: in Section 2, we briefly review basic concepts of MLP and PSNN. In Section 3, we described the dataset used in this study and the model identification for temperature forecasting in Batu Pahat, Malaysia. Section 4 and Section 5 presents the performance comparison metrics in this study and training of the ANN, respectively while Section 6 presents the result by comparing the performance of both networks for one-step-ahead forecasting. The conclusions are discussed in Section 7.

2 Model Description

In this section, we briefly review the concepts of MLP and PSNN.

2.1 Multilayer Perceptron

The MLP, which is a key development in the field of machine learning emulates the biological nervous system's by distributing computations to processing units termed neurons to perform computational tasks [18], [19]. The MLP has the ability to learn from input-output pairs [20] and is capable of solving highly nonlinear problems [21]. The neurons are grouped in layers and adjacent layers that are connected through weights [19]. The MLP adaptively change their synaptic weights through the process of learning. A typical MLP usually has three layers of neurons: input layers, summing layers and output layers [16]. Each hidden and output neurons take linear combination of all values of the previous layer and transformed it with the sigmoid function $1/\left(1+e^{-x}\right)$. The result then is given to the next layer [16] to produce output which is based upon theorem of weighted values passed to them [22]. The sigmoid function acts as a squashing function that prevents accelerating growth throughout the network [23], [24]. The weights of the linear combination are adjusted in the course of training for achieving the desired input-output relation of the network [22] by minimising the error function. It is noted that the MLP has successfully been applied in many applications involving pattern recognition [25], signal processing [26], [27], and classification [28]. However, the ordinary MLP is extremely slow due to its multilayer architecture especially when the summing layer increases. MLP converge very slow in typical situations especially when dealing with complex and nonlinear problems. They are also effectively do not scale the network size [15].

2.2 Pi-Sigma Neural Network

In an effort to overcome the limitations of the ordinary MLP, PSNN that can be described as a multilayer higher order neural network (HONN) has turned researchers' attention. PSNN consists of a single layer of tuneable weights [17], leads MLP in terms of weights and nodes which make the convergence analysis of the learning rules for the PSNN more accurate and tractable [17], [29]. Fig. 1 shows the architecture of k-th Order PSNN which consists of two layers; the product unit and the summing unit layers.

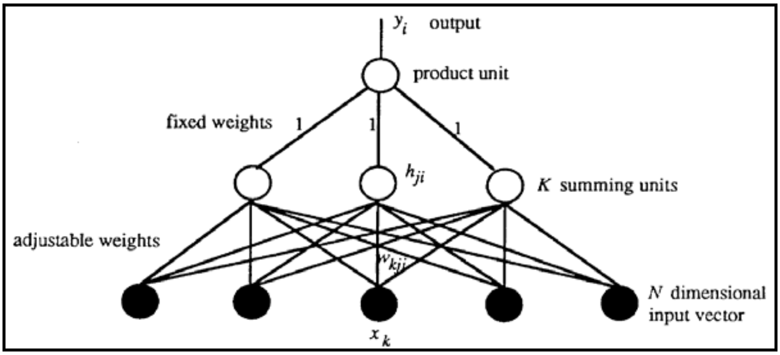


Fig. 1. Structure of k-th Order PSNN.

Input x is an N dimensional vector and x_k is the k-th component of x. The weighted inputs are fed to a layer of K linear summing units; h_{ji} is the output if the j-th summing units for the i-th output y_i, viz:

$$h_{ji} = \sum_k w_{kji} x_k + \theta_{ji}, \text{ and } y_i = \sigma\left(\prod_j h_{ji}\right). \tag{1}$$

where w_{kji} and θ_{ji} are adjustable coefficients, and σ is the nonlinear transfer function [17]. The number of the summing units in PSNN reflects the network order. By using an additional summing unit, it will increase the network's order by 1. In this study, both hidden nodes and higher order terms for MLP and PSNN, respectively are set between 2 and 5. In PSNN, weights from summing layer to the output layer are fixed to unity, resulting to a reduction in the number of tuneable weights. Therefore, it can reduce the training time. Sigmoid and linear functions are adopted in the summing layer and output layer, respectively. The applicability of this network was successfully applied for function approximation [15], pattern recognition [15], classification [17], [30], and so forth. Compared to other HONN, Shin and Ghosh [17] argued that PSNN can contribute to maintain the high learning capabilities of HONN, whilst outperformed the ordinary MLP for similar performance levels, and over a broad class of problems [29].

3 Data Description

We are considering homogenous input data; a series of historical temperature data ranging from 2005 to 2009 as inputs to the network. The data was obtained from Malaysian Meteorological Department (MMD), Malaysia. The selection of ANN input variables are generally based on priori knowledge of the problem under consideration. Therefore, in this study, we use a trial-and-error procedure to determine the input which lies between 4 and 8. As for forecasting horizon, we choose one-step-ahead prediction since the main target is to predict the upcoming measure of daily temperature. To eliminate the non-linearities, the data then is normalised between upper and lower bounds of network's transfer function in the range of $[0.2, 0.8]$.

Each of data series is partitioned into three parts. They are segregated into 50% of training, 25% of testing and 25% of validation. The training set serves the model for training purpose and the testing set is used to evaluate the network performances [21]. Training involves the adjustment of the weights so that the ANN is capable to predict the value assigned to each member of the training set. During the training, the actual and predicted output are compared and weights are adjusted by using the gradient descent rule [31]. The ANN is trained to a satisfactory level with proper features and architectures that able to perform better in prediction [7]. Meanwhile, the validation set has dual-function: 1) to implement an early stopping in order to prevent the training data from overfitting and 2) to select the best predictions from a number of ANN's simulations. On the other hand, the testing set is for generalisation purpose [21], which means producing appropriate outputs for those input samples which were not encountered during training [7].

4 Evaluation of Model Performances

In this study, the Mean Squared Error (MSE), Signal to Noise Ratio (SNR) [29], Mean Absolute Error (MAE) [32] and Normalised Mean Squared Error (NMSE) [29] were used to evaluate the performance of the network models, which are, expressed by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - P_i^*\right)^2 \tag{2}$$

$$\begin{aligned} SNR &= 10 * \lg(sigma) \\ sigma &= \frac{m^2 * n}{SSE} \\ SSE &= \sum_{i=1}^{n}\left(P_i - P_i^*\right) \\ m &= \max(P) \end{aligned} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|P_i - P_i^*\right| \tag{4}$$

$$NMSE = \frac{1}{\sigma^2 n}\sum_{i=1}^{n}\left(P_i - P_i^*\right)^2$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(P_i - P_i^*\right)^2 \tag{5}$$

$$P^* = \sum_{i=1}^{n} P_i$$

where n is the total number of data patterns, P_i and P_i^* represent the actual and predicted output value [29], [32]. In addition, we also consider the number of iterations and CPU Time during the training process.

5 Training of the Artificial Neural Network

The identification of ANN structure is to optimise the number of hidden nodes/higher order terms, h in the summing layer with the known model inputs and output by using training data. Theoretically, Cybernko [24] proved that one-summing-layer using sigmoid transfer function is sufficient enough for the network to memorise patterns and to approximate continuous functions in order to find the local characteristics of the variables examined. Consequently, one summing layer is perfectly enough if no constraints are placed on the number of hidden neurons to model any solution function of practical interest [33], [34]. Additional summing layers might lead to enormous complexity and overfitting problem. Therefore, the identified architecture of ANN models in this study is a three-layered network, which comprise of an input layer, a summing layer and an output layer. The learning rate, η governs the rate at which the weights are allowed to change at any given presentation. Higher η will accelerate the convergence, however, it may lead to oscillations in weight corrections during training, which could expose the integrity of the network, and may cause the learning process to fail [35]. To avoid oscillations and to improve convergence, a smoothing factor, α (momentum) is generally used to ignore small features in the error surface, and therefore, will improve the network performance [35], [36]. The optimal size h of the summing layer is found by systematically increasing the number of hidden neurons and higher order terms for MLP and PSNN, respectively from 2 to 5 until the network performance on the validation is set no longer improves significantly.

Among various kinds of training algorithms, the popular and extensively tested backpropagation method [31], [37] was chosen. In order to minimise the error function at each iteration, the weights are adjusted to ensure the stability and differentiability of the error function, and it is important to ensure the non-existence of regions which the error function is completely flat. As the sigmoid always has a positive derivative, the slope of the error function provides a greater or lesser descent direction which can be followed. The combination of weights which minimises the error function is considered to be a solution of the learning problem [38].

6 Results and Discussions

We have implemented the models using MATLAB 7.10.0 (R2010a) on Pentium® Core ™^{2} Quad CPU. In the present study, the average results of 10 simulations have been conducted on the same training, testing and validation sets. The reason to initialise weights with small values between $(0,1)$ is to prevent saturation for all patterns and the insensitivity to the training process. On the other hand, if the initial weights are set too small, training will tend to start very slow. Early stopping is use as one of the stopping criteria. The stopping criteria are taking under consideration during the learning process. If the validation error continuously increased several times, the network training will be terminated. The weights at the minimum validation error are stored for network generalisation purpose. In the testing phase, this set of weights from the lowest validation error that was monitored during training phase is used. The target error is set to 0.0001 and the maximum epochs to 3000. Table 1 and Table 2 show the simulation results for the PSNN and the MLP, respectively.

Table 1. Average Result of PSNN for One-Step-Ahead Prediction

No. of Input Nodes	Network Order	MAE	NMSE	SNR	MSE Training	MSE Testing	Epoch	CPU Time
4	2	0.0635	0.7791	18.7104	0.0062	0.0065	1211.8	108.80
	3	0.0635	0.7792	18.7097	0.0062	0.0065	1302.9	233.68
	4	0.0635	0.7797	18.7071	0.0062	0.0065	1315.3	366.96
	5	0.0636	0.7822	18.6935	0.0062	0.0066	1201.0	494.56
5	2	0.0631	0.7768	18.7234	0.0061	0.0065	1222.5	112.98
	3	0.0632	0.7769	18.7226	0.0061	0.0065	1221.6	234.29
	4	0.0632	0.7775	18.7193	0.0061	0.0065	1149.9	355.84
	5	0.0633	0.7806	18.7023	0.0062	0.0065	916.9	458.76
6	2	0.0631	0.7758	18.7289	0.0061	0.0065	961.3	92.37
	3	0.0632	0.7770	18.7222	0.0061	0.0065	852.5	180.98
	4	0.0632	0.7775	18.7192	0.0061	0.0065	1155.6	310.29
	5	0.0635	0.7832	18.6880	0.0062	0.0066	948.0	423.68
7	2	0.0630	0.7726	18.7470	0.0061	0.0065	1064.0	106.45
	3	0.0630	0.7733	18.7432	0.0061	0.0065	911.6	205.83
	4	0.0631	0.7760	18.7277	0.0061	0.0065	922.1	314.09
	5	0.0632	0.7766	18.7244	0.0061	0.0065	1072.2	449.68
8	2	**0.0626**	**0.7674**	**18.7688**	**0.0061**	**0.0064**	**719.3**	**74.16**
	3	0.0627	0.7677	18.7671	0.0061	0.0064	877.0	173.73
	4	0.0627	0.7693	18.7579	0.0061	0.0065	861.4	279.71
	5	0.0628	0.7694	18.7574	0.0061	0.0065	970.9	408.92

Table 2. Average Result of MLP for One-Step-Ahead Prediction

No. of Input Nodes	No. of Hidden Nodes	MAE	NMSE	SNR	MSE Training	MSE Testing	Epoch	CPU Time
4	2	0.0636	0.7815	18.6971	0.0062	0.0065	2849.9	324.26
	3	0.0637	0.7831	18.6881	0.0062	0.0066	2468.2	606.15
	4	0.0638	0.7825	18.6918	0.0062	0.0066	2794.2	925.29
	5	0.0638	0.7827	18.6903	0.0062	0.0066	2760.9	1242.21
5	2	0.0632	0.7803	18.7037	0.0062	0.0065	**2028.6**	323.10
	3	0.0633	0.7792	18.7097	0.0062	0.0065	2451.8	602.07
	4	0.0634	0.7789	18.7114	0.0062	0.0065	2678.1	906.99
	5	0.0635	0.7792	18.7097	0.0062	0.0065	2565.8	1200.46
6	2	0.0631	0.7750	18.7335	0.0061	0.0065	2915.1	331.88
	3	0.0633	0.7763	18.7261	0.0062	0.0065	2837.2	655.24
	4	0.0634	0.7776	18.7188	0.0062	0.0065	2652.4	958.81
	5	0.0634	0.7783	18.7152	0.0061	0.0065	2590.8	1256.28
7	2	0.0631	0.7742	**18.7378**	0.0061	0.0065	2951.2	337.62
	3	0.0632	0.7780	18.7164	0.0061	0.0065	2566.6	632.05
	4	0.0632	0.7771	18.7217	0.0061	0.0065	2796.4	1186.71
	5	0.0633	0.7786	18.7131	0.0061	0.0065	2770.0	1897.30
8	2	**0.0629**	**0.7734**	18.7350	**0.0061**	**0.0065**	2684.8	**306.90**
	3	0.0630	0.7749	18.7268	0.0061	0.0065	2647.2	610.67
	4	0.0630	0.7747	18.7278	0.0061	0.0065	2557.1	905.12
	5	0.0631	0.7753	18.7242	0.0061	0.0065	2774.6	1225.88

As it can be noticed, Table 1 shows the results for temperature prediction using PSNN. At the first sight, it would be appear that PSNN reasonably well at temperature forecasting by demonstrating the best results using all measuring criteria with input node equal to 8 and the 2^{nd} order of PSNN. Shortly, it can be said that PSNN with architecture 8-2-1 shows the best results in predicting the temperature. It is same goes to the MLP, which signifies the MLP with the same network architecture to be the best model for temperature forecasting (refer to Table 2). When forecasting the temperature, it can be perceived that PSNN converge faster compared to the MLP for all 810 network architectures that have been trained and tested. This proves that by reducing the number of tuneable weights in the network model, the network can tremendously shorten the training time.

Table 3. Best Results for the PSNN and MLP

Network	MAE	NMSE	SNR	MSE Training	MSE Testing	Epoch	CPU Time
PSNN	0.0626	0.7674	18.7688	0.00612	0.0064	719.3	74.16
MLP	0.0629	0.7734	18.7350	0.00615	0.0065	2684.8	306.90

Table 3 shows the best simulation results for PSNN and MLP. As depicted in the table, it can be seen that PSNN leads MLP by 0.48% for MAE, 0.78% for NMSE, 0.18% for SNR, 0.48% for MSE Training, 1.54% for MSE Testing, 73.21% for Epoch and 75.84% for CPU Time. By considering the MAE, it shows how close forecasts that have been made by PSNN are to the actual output in analysing the results. Concurrently, PSNN beat out the MLP by having smaller percentage of NMSE. Therefore, it can be said that PSNN outperformed MLP in terms of bias and scatter between the predicted and the actual values. In training samples, the MSE for PSNN had shown an improvement by leading the MLP merely to 3×10^{5} as for training, 1×10^{4}. While showing huge comparison in epochs and CPU Time, it can be wrapped up that PSNN not only forecast with minimum error but also can converge faster compared to MLP. Fig. 2 and Fig. 3 represent the best forecast made by PSNN and MLP on temperature, correspondingly. The blue line represents the actual values while the red and black line refers to the predicted values.

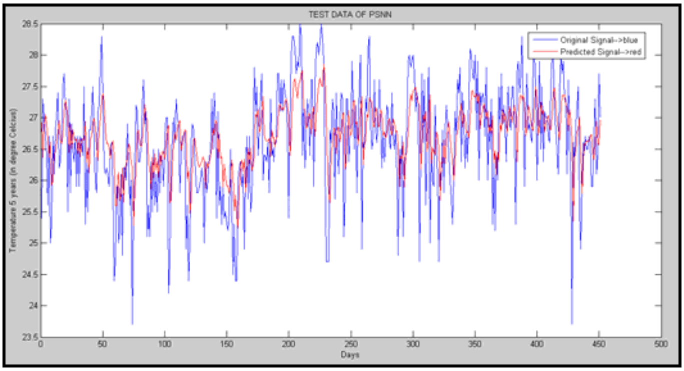


Fig. 2. Temperature Forecast made by PSNN on the Testing Set

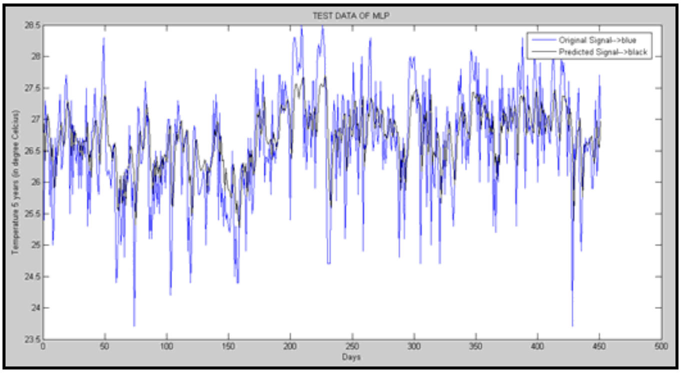


Fig. 3. Temperature Forecast made by MLP on the Testing Set

Graphically, it is verified that PSNN has the ability to perform an input-output mapping of temperature data as well as better performance when compared to MLP for all measurement criteria. More specifically, the PSNN can approximate arbitrarily closely to the actual values. The better performance of temperature forecasting is allocated based on the vigour properties it contains. Hence, it can be seen that PSNN reached higher value of SNR which shows the network can track the signal better than MLP. The thrifty representation of higher order terms in PSNN facilitates the network to model effectively. Accordingly, when compared the MLP with PSNN in terms of CPU Time, it gives an enormous comparison. This is due to the fact that the properties of MLP which has problems in dealing with large amount of training data and require longer time to complete the learning process. While showing a small difference between both models, some might said that it is insignificant to implement PSNN as an alternative for MLP, whilst MLP still can gives an outstanding result, without considering the epochs and the CPU Time. However, it has to be mentioned that no one had ever compared the PSNN with the MLP for temperature forecasting for years. Therefore, the comparison that has been made for both models are still significant to the problem under study, even though the deviation is relatively small.

7 Conclusion

Two ANN models, PSNN and MLP were created and simulated for temperature forecasting. The performances of both models are validated in two ways: (a) the minimum error that can be reached in both training and testing; and (b) the speed of convergence measured in number of iterations and CPU time. From the extensive simulation results, it can be simplified that PSNN with architecture 8-2-1, with learning rate 0.1, momentum 0.2, provides better prediction compared to the MLP. Hence, it can be concluded that PSNN are capable of modelling a temperature forecast for one-step-ahead prediction. As for future works, we are considering on using the temperature historical data for two-step-ahead and expanded to heterogeneous temperature parameters for the robustness of the network model.

Acknowledgments. The authors would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research under the Postgraduate Incentive Research Grant.

References

1. Malaysia Tourism Guide, `http://www.malaysia-tourism-guide.com`
2. MalaysiaVacancyGuide, `http://www.malaysiavacationguide.com`
3. Childs, P.R.N.: Temperature. In: (eds.) Practical Temperature Measurement, pp. 1–15. Butterworth-Heinemann, Butterworths (2001)
4. Ibrahim, D.: Temperature and its Measurement. In: (eds.) Microcontroller Based Temperature Monitoring & Control, Newnes, pp. 55–61 (2002)
5. Barry, R., Chorley, R.: Atmosphere, Weather, and Climate. Methuen (1982)
6. Lorenc, A.C.: Analysis Methods for Numerical Weather Prediction. Quarterly Journal of the Royal Meteorological Society 112, 1177–1194 (1986)
7. Paras, S., Mathur, A.: Kumar and M. Chandra: A Feature Based Neural Network Model for Weather Forecasting. In: Proceedings of World Academy of Science, Engineering and Technology, pp. 67–73 (2007)

8. Chang, F.-J., Chang, L.-C., Kao, H.-S., Wu, G.-R.: Assessing the effort of meteorological variables for evaporation estimation by self-organizing map neural network. Journal of Hydrology 384, 118–129 (2010)
9. Hayati, M., Mohebi, Z.: Application of Artificial Neural Networks for Temperature Forecasting. World Academy of Science, Engineering and Technology 28, 275–279 (2007)
10. Lee, L.-W., Wang, L.-H., Chen, S.-M.: Temperature Prediction and TAIFEX Forecasting based on High-Order Fuzzy Logical Relationships and Genetic Simulated Annealing Techniques. Expert Systems with Applications 34, 328–336 (2008)
11. Smith, B.A., Hoogenboom, G., McClendon, R.W.: Artificial Neural Networks for Automated Year-Round Temperature Prediction. Computers and Electronics in Agriculture 68, 52–61 (2009)
12. Radhika, Y., Shashi, M.: Atmospheric Temperature Prediction using Support Vector Machines. International Journal of Computer Theory and Engineering 1, 55–58 (2009)
13. Wang, N.-Y., Chen, S.-M.: Temperature Prediction and TAIFEX Forecasting based on Automatic Clustering Techniques and Two-Factors High-Order Fuzzy Time Series. Expert Systems with Applications 36, 2143–2154 (2009)
14. Baboo, S.S., Shereef, I.K.: An Efficient Weather Forecasting System using Artificial Neural Network. International Journal of Environmental Science and Development 1, 321–326 (2010)
15. Ghazali, R., Jumeily, D.: Application of Pi-Sigma Neural Networks and Ridge Polynomial Neural Networks to Financial Time Series Prediction. In: Zhang, M. (ed.) Artificial Higher Order Neural Networks for Economics and Business, Information Science Reference, pp. 271–293 (2009)
16. Zhang, G., Patuwo, B.E., Hu, M.Y.: Forecasting with Artificial Neural Networks: The State of the Art. International Journal of Forecasting 14, 35–62 (1998)
17. Shin, Y., Ghosh, J.: The Pi-Sigma Networks: An Efficient Higher-Order Neural Network for Pattern Classification and Function Approximation. In: Proceedings of International Joint Conference on Neural Networks, Seattle, WA, USA, pp. 13–18 (1991)
18. Kröse, B., van der Smagt, P.: An Introduction to Neural Networks. The University of Amsterdam, Netherlands (1996)
19. Brath, A., Montanari, A., Toth, E.: Neural Networks and Non-Parametric Methods for Improving Real-Time Flood Forecasting through Conceptual Hydrological Models. Hydrology and Earth System Sciences 6, 627–640 (2002)
20. Güldal, V., Tongal, H.: Comparison of Recurrent Neural Network, Adaptive Neuro-Fuzzy Inference System and Stochastic Models in Eğirdir Lake Level Forecasting. Water Resources Management 24, 105–128 (2010)
21. Shrestha, R.R., Theobald, S., Nestmann, F.: Simulation of Flood Flow in a River System using Artificial Neural Networks. Hydrology and Earth System Sciences 9, 313–321 (2005)
22. Rumbayan, M., Nagasaka, K.: Estimation of Daily Global Solar Radiation in Indonesia with Artificial Neural Network (ANN) Method. In: Proceedings of International Conference on Advanced Science, Engineering and Information Technology (ISC 2011), pp. 190–103 (2011); Equitorial Bangi-Putrajaya Hotel (2011)
23. Fulcher, J., Jain, L., de Garis, H.: Artificial Brains: An Evolved Neural Net Module Approach. In: (eds.) Computational Intelligence: A Compendium, vol. 115, pp. 797–848. Springer, Berlin (2008)
24. Cybenko, G.: Approximation by Superpositions of a Sigmoidal Function. Signals Systems 2, 14 (1989)

25. Isa, N.A.M., Mamat, W.M.F.W.: Clustered-Hybrid Multilayer Perceptron Network for Pattern Recognition Application. Applied Soft Computing 11, 1457–1466 (2011)
26. Nielsen, J.L.G., Holmgaard, S., Ning, J., Englehart, K., Farina, D., Parker, P.: Enhanced EMG Signal Processing for Simultaneous and Proportional Myoelectric Control. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2009, pp. 4335–4338 (2009)
27. Li, H., Adali, T.: Complex-Valued Adaptive Signal Processing using Nonlinear Functions. EURASIP Journal on Advances in Signal Processing, 1–9 (2008)
28. Yuan-Pin, L., Chi-Hong, W., Tien-Lin, W., Shyh-Kang, J., Jyh-Horng, C.: Multilayer Perceptron for EEG Signal Classification during Listening to Emotional Music. In: TENCON 2007 - 2007 IEEE Region 10 Conference, pp. 1–3 (2007)
29. Ghazali, R., Hussain, A., El-Dereby, W.: Application of Ridge Polynomial Neural Networks to Financial Time Series Prediction. In: International Joint Conference on Neural Networks (IJCNN 2006), Vancouver, BC, pp. 913–920 (2006)
30. Song, G.: Visual Cryptography Scheme Using Pi-Sigma Neural Networks. In: International Symposium on Information Science and Engineering, pp. 679–682 (2008)
31. Popescu, M.-C., Balas, V., Olaru, O., Mastorakis, N.: The Backpropagation Algorithm Functions for the Multilayer Perceptron. In: Proceedings of the 11th WSEAS International Conference on Sustainability in Science Engineering, pp. 32–37. WSEAS Press (2009)
32. Valverde Ramírez, M.C., de Campos Velho, H.F., Ferreira, N.J.: Artificial Neural Network Technique for Rainfall Forecasting Applied to the São Paulo Region. Journal of Hydrology 301, 146–162 (2005)
33. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Network 2, 359–366 (1989)
34. Hecht-Nielsen, R.: Theory of the Back-propagation Neural Network. In: Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 593–605. IEEE, Washington, DC (1989)
35. Al-Jabri, K.S., Al-Alawi, S.M.: An Advanced ANN Model for Predicting the Rotational Behaviour of Semi-rigid Composite Joints in Fire Using the Back-Propagation Paradigm. International Journal of Steel Structures 10, 337–347 (2010)
36. Huang, B.Q., Rashid, R., Kechadi, M.-T.: Multi-Context Recurrent Neural Network for Time Series Applications. International Journal of Computer Intelligence 3, 45–54 (2007)
37. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Representations by Back-Propagating Errors. Nature 323, 533–536 (1986)
38. Rojas, R.: The Backpropagation Algorithm. Springer, Berlin (1996)

Revisiting Protocol for Privacy Preserving Sharing Distributed Data: A Review with Recent Results

Ahmed HajYasien

IT Department
University College of Bahrain
ayasien@ucb.edu.bh

Abstract. In this paper, we give a review of some recent results concerning the study of an efficient protocol that allows parties to share data in a private way with no restrictions and without loss of accuracy. Privacy policies might discourage people who would otherwise participate in a joint contribution to a data mining task. Previously, we proposed a method that has the immediate application on horizontally partitioned databases which can be brought together and made public without disclosing the source/owner of each record. We also showed an additional benefit that we can apply our protocol to privately share patterns in the form of association rules. We performed more experiments to show that our protocol is more efficient than previous protocols. Aside from the above, we propose a new categorization for the privacy preserving data mining field.

Keywords: Privacy preserving data mining, secure multi-party computation, data sanitization, data privacy, sensitive knowledge, association rule-mining.

1 Introduction

Information communication technology (ICT) has made this world very competitive with more and more privacy breaches. In their struggle to preserve client's rights, to approach new clients or even to enhance services and decision making, data owners need to share their data for the common good. This philosophy collides with privacy rights and the need for security. Privacy concerns have been influencing data owners and preventing them from achieving the maximum benefit of data sharing. Data owners usually sanitize their data and try to block as many inference channels as possible to prevent other parties from concluding what they consider sensitive. Data sanitization is defined as the process of making sensitive information in non-production databases safe for wider visibility [1]. However, sanitized databases are presumed secure and useful for data mining, in particular, for extracting association rules.

Oliveira et al. [2] classified the existing sanitizing algorithms into two major classes: data-sharing techniques and pattern-sharing techniques (see Figure 1).

I Data-sharing techniques communicate data to other parties without analysis or summarization with data mining or statistical techniques. Under this approach, researchers proposed algorithms that change databases and produce distorted databases in order to hide sensitive data. Data-sharing techniques are, in themselves, categorized

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 542–555, 2011.

as follows: First, (item restriction)-based algorithms. In this class, the methods [3], [4], [5], [6] reduce either the support or confidence to a safe zone (below a given privacy sup- port threshold) by deleting transactions or items from a database to hide sensitive rules that can be derived from that database. Second, (item addition)-based: algorithms. This group of methods [3] adds imaginary items to the existing transactions. Usually the addition is performed to items in the antecedent part of the rule. As a result, the confidence of such a rule is reduced and enters the safe zone. The problem with this approach is that the addition of new items will create new rules and parties could share untrue knowledge (sets of items that are not frequent itemsets appear as such). Third, (item obfuscation)-based: algorithms. The algorithms [7] replace some items with a question mark in some transactions to avoid the exposure of sensitive rules. Unlike the (item addition)-based, the (item obfuscation)-based, approach saves parties from sharing false rules.

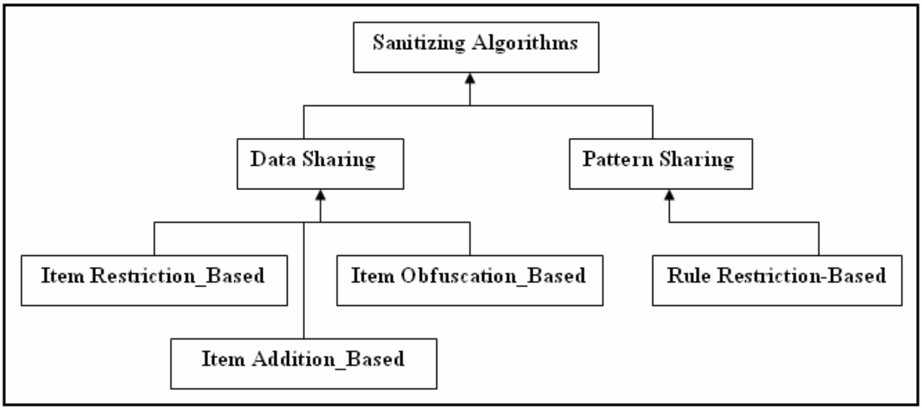


Fig. 1. Taxonomy of Sanitizing Algorithms

II This second major class are pattern-sharing techniques, where the sanitizing algorithms act on the rules mined from a database, instead of the data it- self. The existing solutions either remove all sensitive rules before the sharing process [2] (such solutions have the advantage that two or more parties can apply them) or share all the rules in a pool where no party is able to identify or learn anything about the links between individual data owned by other parties and their owners [8] (such solutions have the disadvantage that they can be applied only to three or more parties).

We believe that the above categorization for where the privacy preserving data mining occurs is not very clear.

We present a new categorization that is based on “classification by the where”. We believe our classification is general, comprehensive and gives better understanding to the field of PPDM in terms of laying each problem under the right category. The new classification is as follows: PPDM can be attempted at three levels as shown in Figure 2. The first level is raw data or databases where transactions reside. The second level is data mining algorithms and techniques that ensure privacy. The third level is the output of different data mining algorithms and techniques.

In our opinion, privacy preserving data mining occurs in two dimensions inside each level of three levels mentioned above as follows:

- Individuals: This dimension involves implementing a privacy preserving data mining technique to protect the privacy of one individual or more whose their data is going to be published to the public. An example of this dimension is patients' records or the census.
- PPDMSMC (Privacy Preserving Data Mining in Secure Multiparty Computation): This dimension involves protecting the privacy of two or more parties who want to perform a data mining task on the union of their private data. An example of this dimension is two parties who want to cluster the union of their private data.

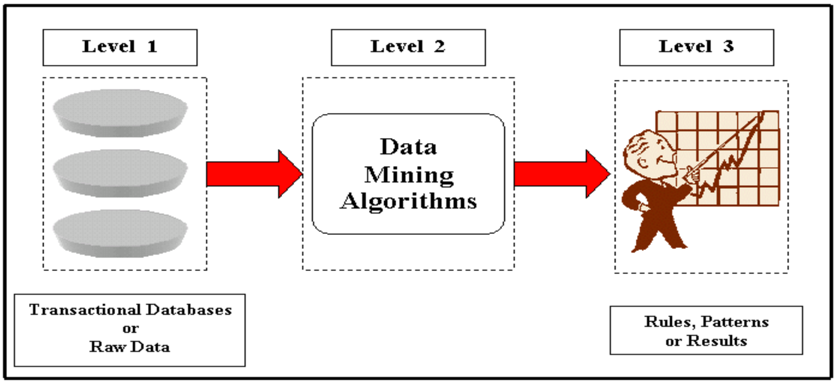

Level 1
Level 2
Level 3
Data Mining Algorithms
Transactional Databases or Raw Data
Rules, Patterns or Results

Fig. 2. Three Major Levels Where Privacy Preserving Data Mining Can Be Attempted

2 Motivation

As of May 2009, the size of the world's total digital content has been roughly estimated to be 500 exabytes [9]. An Exabyte is a unit of information or computer storage equal to 10^{18} bytes or 1 billion gigabytes. Data mining can be a powerful means of extracting useful information from these data. As more and more digital data becomes available, the potential for misuse of data mining grows. Different organizations have different reasons for considering specific rows or patterns in their huge databases as sensitive. They can restrict what to expose to only what is necessary! But who can decide what is necessary and what is not? There are scenarios where organizations from the medical sector or the government sector are willing to share their databases as one database or their patterns as one pool of patterns if the transactions or patterns are shuffled and no transaction or pattern can be linked to its owner.

These organizations are aware that they will lose if they just hide the data and do not implement privacy practices to share it for the common good. Recently, many countries have promulgated new privacy legislation. Most of these laws incorporate rules governing collection, use, store, share and distribution of personally identifiable in- formation. It is up to an organization to ensure that data processing operations respect any legislative requirements. These organizations that do not respect the legislative requirements can harm themselves or the others by exposing sensitive knowledge and can be sued. Further, client/organization relationships are built on trust. Organizations that demonstrate and apply good privacy practices can build trust.

3 Related Work

Researchers have proposed algorithms that perturb data to allow public disclosure or for a privacy preserving data mining in secure multi-party computation task (PPDMSMC) (explained in Section 4.2) [10], [11], [12], [13], [14], [15]. The balance between privacy and accuracy on data- perturbation techniques depends on modifying the data in a way that no party can reconstruct data of any individual transaction but the overall mining results are still valid and close to the exact ones. In other words, the more the distortion to block more inference channels, the less accurate the results will be. In general, it has been demonstrated that in many cases random data distortion preserves very little privacy [16]. The PPDMSMC approach uses cryptographic tools to the problem of computing a data-mining task from distributed data sets, while keeping local data private [17], [18], [19], [20], [21], [22]. These tools allow parties to analyze their data and achieve results without any disclosure of the actual data. Murat et al. [8] proposed a method that incorporates cryptographic techniques and show frequent itemsets, of three or more parties, as one set where each party can recognize its itemsets but can not link any of the other itemsets to their owners. This particular method uses commutative encryption which could be very expensive if we have large number of parties. Another shortcoming of this method is that when a mining task is performed on the joint data, the results also are published to all parties, and parties are not free of choosing the mining methods or parameters. This might be convenient, if all parties need to perform one or more tasks on the data and all parties agree to share the analysis algorithms. On the other hand, par- ties might wish to have access to the data as a whole and perform private analysis and keep the results private. Our protocol offers exactly this advantage and as we mentioned earlier, there are no limitations on the analysis that each party can perform privately to the shared data.

4 Preliminaries

In the following we will cover some basic concepts that will help making this paper clear.

4.1 Vertical vs. Horizontal Distribution

With vertically partitioned data, each party collects different attributes for the same objects. In other words, attributes or columns will be distributed among database owners. For example, patients have attributes with hospitals different from attributes with insurance companies.

With horizontally partitioned data, each party collects data about the same attributes for objects. In other words, transactions or rows will be distributed among database owners. For example, hospitals that collect similar data about different diseases for different patients.

If the data is distributed vertically, then unique at- tributes that appear in a pattern or a transaction can be linked to the owner. In this paper we assume horizontal distribution of data.

4.2 Security Multi-Party Computation

The problem of secure multi-party computation is as follows: N parties, $P_0, \dots, P_n$ wish to evaluate a function $F(x_1, \dots, x_n)$, where x_i is a secret value provided by P_i. The goal is to preserve the privacy of the each party's in- puts and guarantee the correctness of the computation. This problem is trivial if we add a trusted third party *T to* the computation. Simply, T collects all the inputs from the parties, computes the function F, and announces the result. If the function F to be evaluated is a data mining task, we call this privacy preserving data mining in secure multi-party computation (PPDMSMC).

It is difficult to agree on a trusted party in the industrial sector. The algorithms proposed to solve PPDMSMC problems usually assume no trusted party, but assume a semi-honest model. The semi-honest model is a more realistic abstraction of how parties would engage and participate in a collective computation while preserving each the privacy of their data.

4.3 Two Models

In the study of secure multi-party computation, one of two models is usually assumed: the malicious model and the semi-honest model.

4.3.1 The Malicious Model

The malicious party is a party who does not follow the protocol properly. The model consists of one or more malicious parties which may attempt to deviate from the protocol in any manner. The malicious party can deviate from the protocol through one of the following possibilities:

- A party may refuse to participate in the protocol when the protocol is first invoked.
- A party may substitute his local input by entering the protocol with an input other than the one provided to them.
- A party may abort prematurely.

4.3.2 The Semi-honest Model

A semi-honest party is one who follows the protocol steps but feels free to deviate in between the steps to gain more knowledge and satisfy an independent agenda of interests [23]. In other words, a semi-honest party follows the protocol step by step and computes what needs to be computed based on the input provided from the other parties, but it can do its own analysis during or after the protocol to compromise privacy/security of other parties. It will not insert false information that will result in failure to compute the data mining result, but will use all the information gained to attempt to infer or discover private values from the data sets of other parties. A definition of the semi- honest model [24] formalizes that whatever a semi-honest party learns from participating in the protocol, this information could be essentially obtained from its inputs and its outputs. In particular, the definition uses a probabilistic functionality f : {0, 1}* x {0, 1}* →{0, 1}* x {0, 1}* computable in polynomial-time. Here, f_1 (x, y) denotes the first element of f (x, y), and says that what the output string is for the first party as a function of the inputs strings of the two parties (and f_2 (x, y) is the respective second component of f (x, y) for the second party). The two-party protocol is denoted by Π. VIEW^{Π} (*x, y*) denotes the view of the first party during an execution of Π on (x, y). Such view consists of (x, r, m_1 , . . . , m_t), where *r* represent the outcome of the first party's internal coin tosses and m_i rep- resents the i^{th} message it has received. Then, Π can privately compute f , with respect to the first party, if there exist probabilistic polynomial time algorithms S_1 such that even if party two provides arbitrary answers during the protocol, the corresponding view for the first party is the out- put of the algorithm S_1 on the input *x* of the first party and the messages received by the first party. The protocol can privately compute f if it can do so with respect to both parties. The theory of this model [24] shows that to compute privately under the semi-hones model is also equivalent to compute privately and securely. Therefore, the discussion of this model assumes parties behaving under the semi-honest model. In the following we explain what is public-key cryptosystem.

4.4 Public-Key Cryptosystems (Asymmetric Ciphers)

A Cipher is an algorithm that is used to encrypt plaintext into ciphertext and vice versa (decryption). Cipher's are said to be divided into two categories: private key and public key. Private Key (symmetric key algorithms) requires a sender to encrypt a plaintext with the key and the receiver to decrypt the ciphertext with the key. We can see, a problem with this method is both parties must have a identical key, and somehow the key must be delivered to the receiving party. Public Key (asymmetric key algorithms) uses two separate keys: a public key and a private key. The pub- lic key is used to encrypt the data and only the private key can decrypt the data. A form of this type of encryption is called RSA (discussed below), and is widely used for se- cured websites that carry sensitive data such as username and passwords, and credit card numbers.

Public-key cryptosystem were invented in the late 1970's, along developments in complexity theory [25],[26]. As a result, Cryptosystems could be developed which would have two keys, a private key and a public key. With the public key, one could encrypt data, and decrypt them with the private key. Thus, the owner of the private key would be the only one who could decrypt the data, but any- one knowing the public key could send them a message in private. Many of the public key systems are also patented by private companies, this also limits their use. For example, the RSA algorithm was patented by MIT in 1983 in the United States of America as (U.S. patent #4,405,829). The patent expired on 21 September 2000.

The RSA algorithm was described in 1977 [27] by Ron Rivest, Adi Shamir and Len Adleman at MIT; the letters RSA are the initials of their surnames. RSA is currently the most important public-key algorithm and the most commonly used. It can be used both for encryption and for digital signatures. RSA computation takes place with integer modulo $n = p * q$, for two large secret primes p and q. To encrypt a message m, it is exponentiated with a small public exponent e. For decryption, the recipient of the ciphertext $c = m^e$ (mod n) computes the multiplicative reverse $d = e^{-1}$ $(mod(p - \mathrm{i}) * (q - \mathrm{i}))$ (we require that e is selected suitably for it to exist) and obtains $c^d = m^{e*d} = m(modn)$. The private key consists of n, p, q, e, d. The public key contains only of n, e. The problem for the attacker is that computing the reverse d of e is assumed to be no easier than factorizing n [25].

The key size (the size of the modulus) should be greater than 1024 bits (i.e. it should be of magnitude 10^{300}) for a reasonable margin of security. Keys of size, say, 2048 bits should give security for decades [28].

Dramatic advances in factoring large integers would make RSA vulnerable, but other attacks against specific variants are also known. Good implementations use redundancy in order to avoid attacks using the multiplicative structure of the ciphertext. RSA is vulnerable to chosen plain-text attacks and hardware and fault attacks. Also, important attacks against very small exponents exist, as well as against partially revealed factorization of the modulus.

The proper implementation of the RSA algorithm with redundancy is well explained in the PKCS standards (see definitions at RSA Laboratories [29]). The RSA algorithm should not be used in plain form. It is recommended that implementations follow the standard as this has also the additional benefit of inter-operability with most major protocols.

5 Statement of the Problem

Let $P = \{P_0, \ldots, P_n\}$ be a set of N parties where $|N| \geq 3$. Each party P_i has a database DB_i . We assume that parties running the protocol are semi-honest. The goal is to share the union of DB_i as one shuffled database $DB_{Comp=} \bigcup_{i=0}^{n} DB_i$ and hide the link between records in DB_{Comp} and their owners.

Our protocol that was presented in [23] employs a public-key cryptosystem algorithm on a horizontally partitioned data among three or more parties. In our protocol, the parties can share the union of their data without the need for an outside trusted party. The information that is hidden is what data records belong to which party. The details of this protocol can be found in [23].

6 Basket Market Application

It may seem that the protocol above is rather elaborate, for the seemingly simple task of bringing the data of all parties together while removing what record (transaction) was contributed by whom. We now show how to apply this protocol to improve on the privacy preserving data mining of association rules.

The task of mining association rules over market basket data [30] is considered a core knowledge discovery activity since it provides a useful mechanism for discovering correlations among items belonging to customer transactions in a market basket database. Let D be the database of transactions and $J = \{J_1, ..., J_n\}$ be the set of items. A transaction T includes one or more items in J *(i.e.,* $T \subseteq J$ *).*An association rule has the form $X \rightarrow Y$, where X and Y are non-empty sets of items (*i.e.* $X \subseteq J$, $Y \subseteq J$) such that $X \cap Y = \emptyset$. A set of items is called an itemset, while X is called the antecedent. The support $sprt_D(x)$ of an item (or itemset) x is the percentage of transactions from D in which that item or itemset occur in the database. In other words, the support s of an association rule $X \rightarrow Y$ is the percentage of transactions T in a database where $X \cup Y \subseteq T$. The confidence or strength c for an association rule $X \rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X. An itemset $X \subseteq J$ is frequent if at least a fraction s of the transaction in a database contains X. Frequent itemsets are important because they are the building block to obtain association rules with a given confidence and support.

The distributed mining of association rules over horizontally partitioned data consists of sites (parties) with homogeneous schema for records that consists of transactions. Obviously we could use our protocol COMBINE_WITHOUT_OWNER [23] to bring all transactions together and then let each party apply an association-rule mining algorithm (Apriori or FP-tree, for example) to extract the association rules. This approach is reasonably secure for some settings, but parties may learn about some transactions on other parties. Ideally, it is desirable to obtain association rules with support and confidence over the entire joint database without any party inspecting other parties transactions [8]. Computing association rules without disclosing individual transactions is possible if we can have some global information. For example, if one knows that 1) ABC is a global frequent itemset, 2) the local support of AB and ABC and 3) the size of each database DB_i, then one can determine if $AB \Rightarrow C$ has the necessary support and confidence [23].

Thus, to compute distributed association rules privately, without releasing any individual transaction, the parties compute individually their frequent itemsets at the desired support. Then, for all those itemsets that are above the desired relative support, the parties use our protocol to share records that consist of the local frequent itemset, and its local support (not transactions). The parties also share the size of their local databases. Note that, because an itemset has global support above the global support at p percent only if at least one party has that itemset as frequent in its database with local support at least p percent, we are sure that the algorithm finds all globally frequent itemsets.

We would like to emphasize two important aspect of our protocol in this application. The first is that we do not require commutative[1] encryption. The second is that we require two exchanges of encrypted data between the parties less than the previous algorithms in [8] for this task as Figure 3 shows. The third, is that we do not require that the parties find first the local frequent itemsets of size 1 in order to find global frequent itemsets of size 1, and then global candidate itemsets of size two (and

[1] Commutative encryption means that if we have two encryption algorithms $E1$ and $E2$, the order of their application is irrelevant; that is $E1(E2(x)) = E2(E1(x))$.

then repeatedly find local frequent itemsets of size *k* in order to share them with others for obtaining global itemsets of size *k* that can then formulate global candidate itemset of size *k* + i).

In our method, each party works locally finding all local frequent itemsets of all sizes. They can use Yao's Millionaire protocol to find the largest size for a frequent local itemset. This party sets the the value *k* and parties use our protocol to share all local and frequent itemset of size *k*. Once global frequent itemsets of size *k* are known, parties can check (using the anti-monotonic property that if an itemset is frequent all its subsets must be frequent) so they do not disclose locally frequent itemsets that have no chance of being globally frequent.

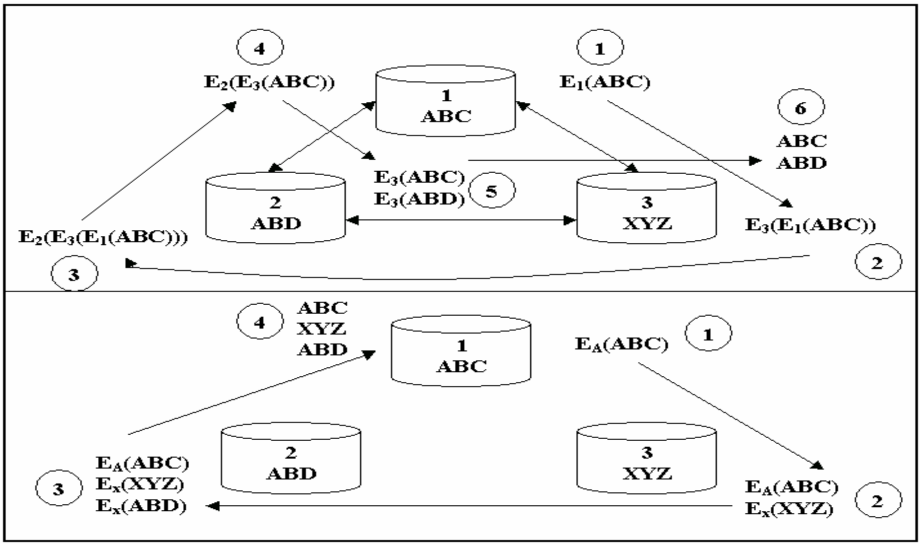


Fig. 3. Steps for Sharing Global Candidate Itemsets Reduced from 6 to 4 Steps

With this last aspect of our protocol we improve the privacy above previous algorithms [8]. Specifically, the contribution here is the sharing process was reduced from 6 steps to 4 steps as Figure 3 shows.

7 Cost of Data Breach

In the past, parties who seek privacy were hesitant to implement database encryption because of the very high cost, complexity, and performance degradation. Recently, with the ever growing risk of data theft and emerging legislative requirements, parties are more willing to compromise efficiency for privacy. The theoretical analysis indicates that the computational complexity of RSA decryption of a single *n* bit block is approximately $O(n^3)$, where *n* denotes both the block length and key length (exponent and modulus). This is because the complexity of multiplication is $O(n^2)$, and the complexity of exponentiation is $O(n)$ when square and multiply is used. The

OpenSSL implementation can be used (with RSA keys) for secure, authenticated communication between different sites [31]. SSL is short for Secure Sockets Layer, a protocol developed by Netscape for transmitting private data via the Internet. The overhead of SSL communication has been found of practical affordability by other researchers [32].

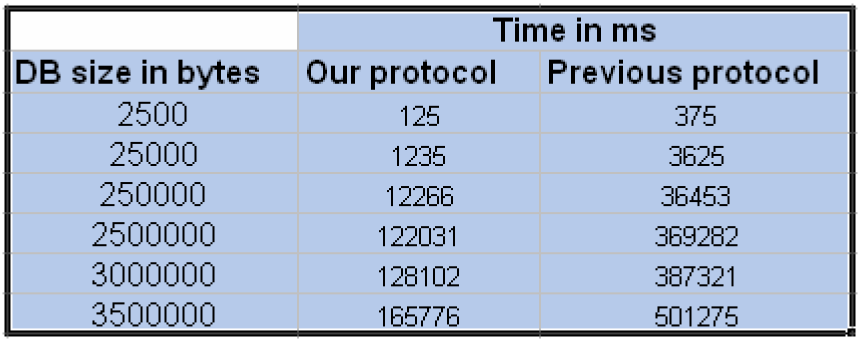
We analyzed the cost of RSA encryption [23] in terms of computation, number of messages, and total size. For this analysis, we implemented RSA in Java to calculate the encryption time of a message of size m = 64 bytes with encryption key of 1024-bits. This time was 0.001462 sec on a 2.4MHz Pentium 4 under Windows. This is perfectly comparable with the practical computational cost suggested by earlier methods [8]. While some regards RSA is too slow for encrypting large volumes of data [33], our implementation is particularly competitive. An evaluation of previous methods [8] suggested that (on distributed association rule mining parameters found in the literature [34]), the total overhead was approximately 800 seconds for databases with 1000 attributes and half a million transactions (on a 700MHz Pentium 3). Our implementation requires 30% of this time (i.e. 234.2 seconds), but on a Pentium 4. In any case, perfectly affordable.

In an extension to the above results achieved in our previous research [23], we performed another set of experiments to compare our protocol to the previous protocol presented in [8]. To achieve the best and unbiased results, we generated random data to create different database sizes from 2,500 bytes to 3,500,000 bytes.

The results in Table 1 shows the comparison of performance between our protocol and the protocol presented in [8]. We can notice that there is a significant difference in the performance of the two protocols where our protocol shows important superiority.

The experiments included the whole steps of each protocol except for shuffling the records which is supposed to take the same amount of time in both of the protocols and thus can be ignored. In other words, the experiments performed included the encryption and decryption of different databases' sizes to compare the performance of our protocol to the other protocol. Figure 4 show that our protocol is significantly faster than the protocol in [8].

Table 1. The comparison of the performance of the two protocols



	Time in ms	
DB size in bytes	Our protocol	Previous protocol
2500	125	375
25000	1235	3625
250000	12266	36453
2500000	122031	369282
3000000	128102	387321
3500000	165776	501275

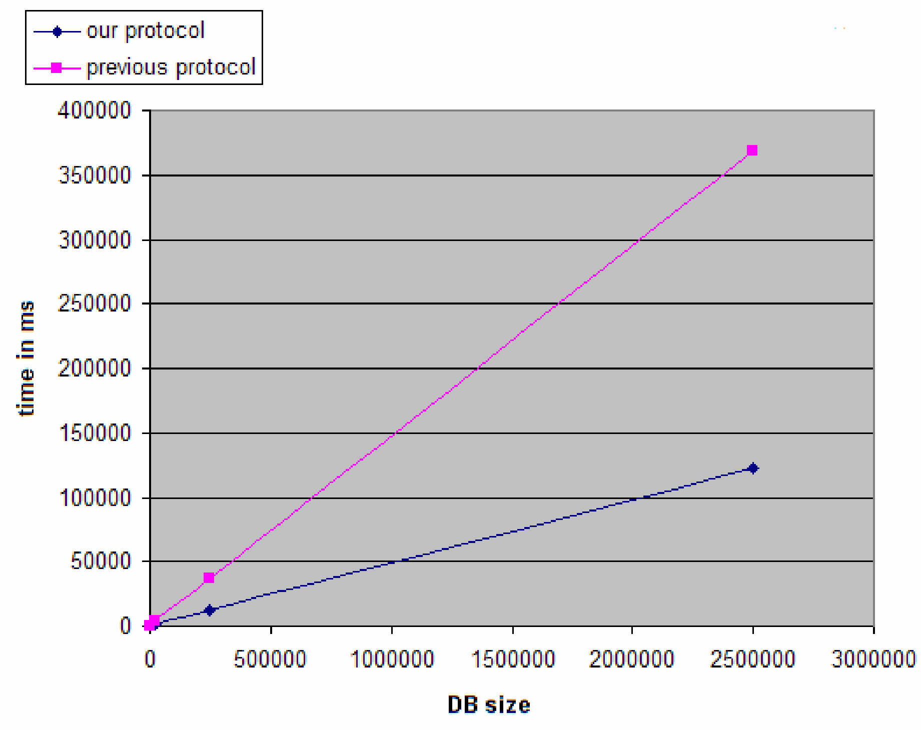


Fig. 4. Our Protocol Compared to Previous Protocol

8 Conclusion

Organizations and people like their data to be strictly protected against any unauthorized access. For them data privacy and security is a priority. At the same time, it might be necessary for them to share data for the sake of getting beneficial results. The problem is how can these individuals or parties compare their data or share it without revealing the actual data to each other. It is also always assumed that the parties who want to compare or share data results do not trust each other and/or compete with each other.

We have presented a new classification for the privacy preserving data mining problems. The new classification is better because individuals or parties who want to protect their data usually look at the privacy preserving tools as a black box with input and output. Usually the focus is not whether the privacy preserving data mining technique is based on cryptography techniques or based on heuristic techniques. What is important is, we want to protect the privacy of our data at Level 1, Level 2 or at Level 3.

In [23], we proposed a flexible and easy-to-implement protocol for privacy preserving data sharing based on a public-key cryptosystem. Through an extensive experiments, we proved that our protocol is efficient in practical and it requires less machinery than previous approaches (where commutative encryption was required).

This protocol ensures that no data can be linked to a specific user. We did not generate an output of a mining task (association rules) but our protocol can be applied either databases or to the output of a mining task (i.e. association rules). Our protocol allows users to conduct private mining analyses without loss of accuracy. A privacy concern of this protocol –in case of sharing actual databases- is that the users get to see the actual data. But previous research has explored that parties are willing to trade off the benefits and costs of sharing sensitive data [35], [36]. The results of this research showed that parties are willing to trade-off privacy concerns for economic benefits. There are few issues that may influence practical usage of the presented protocol. While our protocol is efficient, it may be still a heavy overhead for parties who want to share huge multimedia databases. Also, as we mentioned before; the problem of trusting one party with shuffling the records and publishing the database\or association rules to all parties can be solved with slightly more cost. If there are N parties, each party plays the data distributor with i/N share of the data, and we conduct N rounds.

In summary, with more analysis and extensive experiments, we showed that our protocol not only satisfies the security requirements but also more efficient than the protocol presented in [8]. We showed that the overhead to security is reduced as we do not need commutative encryption , the basket market sharing process was reduced from 6 steps to 4 steps, and the protocol is more secure as we share less local frequent itemsets that may not result in a global frequent itemsets.

References

1. Edgar, D.: Data sanitization techniques. White Papers (2004)
2. Oliveira, S.R.M., Zaïane, O.R., Saygın, Y.: Secure association rule sharing. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 74–85. Springer, Heidelberg (2004)
3. Dasseni, E., Verykios, V.S., Elmagarmid, A.K., Bertino, E.: Hiding association rules by using confidence and support. In: Proc. of the 4th Information Hiding Workshop, Pittsburg, USA, pp. 369–383 (April 2001)
4. Oliveira, S.R.M., Zaiane, O.R.: Privacy preserving frequent itemset mining. In: Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan, pp. 43–54 (December 2002)
5. Oliveira, S.R.M., Zaiane, O.R.: Algorithms for balancing privacy and knowledge discovery in association rule mining. In: Proc. of the 7th In- ternational Database Engineering and Applications Symposium (IDEAS 2003), Hong Kong, pp. 54–63. China (July 2003)
6. Oliveira, S.R.M., Zaiane, O.R.: Protecting sensitive knowledge by data sanitization. In: Proc. of the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourne Florida, USA, pp. 613–616 (November 2003)
7. Saygin, Y., Verykios, V.S., Clifton, C.: Using unknowns to prevent discovery of association rules. SIGMOD Record 30(4), 45–54 (2001)
8. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. In: The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery DMKD 2002 (June 2002)

9. The Guardian (May 18, 2009), http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion (retrieved on April 23, 2010)
10. Agrawal, D., Aggarwal, R.: On the design and quantification of privacy preserving data mining algorithms. In: Proc. of the 20th ACM SIGMOD-SIGACT- SIGART Symposium on Principles of Database Systems, pp. 247–255. ACM Press, New York (2001)
11. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM Press, New York (2000)
12. Du, W.L., Zhan, Z.J.: Using randomized response techniques for privacy-preserving data mining. In: Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 505–510. ACM Press, New York (2003)
13. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In: Proc. of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 211–222. ACM Press, San Diego (2003)
14. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 217–228. ACM Press, New York (2002)
15. Rizvi, S.J., Haritsa, J.R.: Maintaining data privacy in association rule mining. In: Proc. of 28th VLDB Conference (2002)
16. Kargupta, H., Datta, S., Wang, A., Sivakumar, K.: On the privacy preserving properties of ran- dom data perturbation techniques. In: Proc. of the Third IEEE International Conference on Data Mining, ICDM 2003 (2003)
17. Pinkas, B.: Cryptographic techniques for privacy- preserving data mining. In: Proc. of the ACM SIGKDD Explorations, vol. 4, pp. 12–19 (2003)
18. Ambainis, A., Jakobsson, M., Lipmaa, H.: Cryptographic randomized response techniques. In: Bao, F., Deng, R., Zhou, J. (eds.) PKC 2004. LNCS, vol. 2947, pp. 425–438. Springer, Heidelberg (2004)
19. Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: Proc. of ACM SIGMOD. ACM Press, San Diego (2003)
20. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–54. Springer, Heidelberg (2000)
21. Vaidya, J., Clifton, C.: Privacy-preserving k-means clustering over vertically partitioned data. In: Proc. of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 206–215. ACM Press, Washington, D.C (2003)
22. Vaidya, J., Clifton, C.: Privacy preserving nave bayes classifier for vertically partitioned data. In: Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
23. Estivill-Castro, V., HajYasien, A.: Fast Private Association Rule Mining by a Protocol for Securely Sharing Distributed Data. In: Proc. IEEE International Conference on Intelligence and Security Informatics, ISI 2007, New Brunswick, New Jersey, USA, May 23-24, pp. 324–330 (2007)
24. Goldreich, O.: Secure multi-party computation. Working Draft (1998)
25. Menezes, J., Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1996)
26. Schneier, B.: Applied Cryptography. John Wiley and Sons, Chichester (1996)
27. Rivest, R.L., Shamir, A., Adelman, L.M.: A method for obtaining digital signatures and public-key cryptosystems. Technical Report MIT/LCS/TM-82 (1977)
28. Wiener, M.: Performance comparisons of public-key cryptosystems. In: Proc. of the RSA Data Security Conference, San Francisco, USA (January 1998)

29. RSA Laboratories Web Site, `http://www.devx.com/security/link/8206`
30. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of the ACM SIGMOD Conference on Management of Data, Washington D.C., USA, pp. 207–216 (May 1993)
31. OpenSSL, `http://www.openssl.org/`
32. Apostolopoulos, G., Peris, V., Saha, D.: Transport layer security: How much does it really cost? In: INFOCOM: The Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, pp. 717–725 (1999)
33. Tanenbaum, S.: Computer Networks, 3rd edn. Prentice-Hall, Englewood Cliffs (1996)
34. Cheung, W.L., Ng, V., Fu, W.C., Fu, Y.: Efficient mining of association rules of distributed databases. IEEE Transactions Knowledge Data Engineering 8(6), 911–922 (1996)
35. Hann, I.H., Hui, K.L., Lee, T.S., Png, I.P.L.: Online information privacy: Measuring the cost- benefit trade-off. In: Proc. of the Twenty-Third Inter- national Conference on Information Systems(ICIS), Barcelona, Spain (December 2002)
36. Westin, A.: Freebies and privacy: What net users think? Technical report. Opinion Research Corporation 4(3), 26 (1999)

The Performance of Compound Enhancement Algorithm on Abnormality Detection Analysis of Intra-oral Dental Radiograph Images

Siti Arpah Ahmad[1], Mohd Nasir Taib[1], NoorElaiza Abd Khalid[2],
Rohana Ahmad[3], and Haslina Taib[4]

[1] Faculty of Electrical Engineering,
[2] Faculty of Computer and Mathematical Sciences,
[3] Faculty of Dentistry, University Teknology MARA, Shah Alam, Malaysia
[4] School of Dental Sciences, Universiti Sains Malaysia,
Helath Campus, Kubang Kerian, Malaysia
{arpah,elaiza}@tmsk.uitm.edu.my, dr.nasir@ieee.org

Abstract. Dentists look for abnormality in radiograph for determining any diseases that may appear at the apices of the teeth. However poor quality of the radiograph produces weak visual signal that may produce misleading interpretations. Hence the quality of radiograph influence dentists' decision that reflects the success or failure of any suggested treatments. Thus this work aim to analyze the abnormality found in intra-oral dental radiographs by comparing the original images with images that had been enhanced using compound enhancement algorithms (CEA) namely Sharp Adaptive Histogram Equalization (SAHE) and Sharp Contrast adaptive histogram equalization (SCLAHE). Results show that SCLAHE enhanced images provide slight improvement, compared to the original images, in detecting widen periodontal ligament space abnormality

Keywords: Intra-oral dental radiograph; periapical lesion; AHE; CLAHE.

1 Introduction

Radiographic diagnosis influence treatment planning and overall cost of dental health care [1]. Radiographs images are often noisy and low in contrast and sometimes make it difficult to interpret [2]. Studies have proven that image processing techniques can assist dentists to improve diagnosis [3-5]. Contrast enhancement is one of the techniques that are actively being researched to improve the dental radiographs. Even though contrast enhancements are usually built in the software accompanying the x-ray machines, the interactive trial and error adjustment of contrast and brightness is a time-consuming procedure [5]. Thus a more automated and universal contrast enhancement is needed to overcome this problem. This work compares the performance of sharpening function combined with adaptive histogram equalization (SAHE) and sharpening combined with contrast limited adaptive histogram equalization (SCLAHE) with the original image. The tests are limited to assessing the

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 556–566, 2011.

abnormality detection of dedicated pathologies. Comparison and correlation are made between the dentists' perceptions and statistical values such as contrast improvement index (CII), signal to noise ratio (SNR) and root mean square error (RMSE).

2 Research Backgound

Contrast enhancement algorithm has proven to have some impact in improving dental radiographs [2-5]. Contrast is the visual differences between the various black, white and grey shadows exist in an image [6]. Contrast enhancement algorithms are functions that manipulate the brightness intensity of the image by stretching brightness values between dark and bright area [7]. This operation will generate clearer image to the eyes or assist feature extraction processing in computer vision system [8]. Research related to application of several techniques such as high pass and contrast enhancement had been applied to dental radiographs [2-5][9-10]. Some of these techniques are available in the software provided by the x-ray machine vendor such as are Digora for Windows [2], Photoshop 8.0 [3] and Trophy Windows [4]. Algorithm such as Sliding window adaptive histogram equalization (SWAHE) [5] and frequency domain algorithms [9] also provide successful enhancement. High pass filters that have been used are shadow [2] and sharpening [2][10]. Other contrast enhancement variations are adaptive histogram equalization (AHE) [3], bright contrast enhancement [4] and pseudo-colored with brightness-contrast adjustment [4]. Negative or inversion algorithm have been used in [2][4] to test the effect of brightness changes in dark region of images.

The dental anatomical structures that had been investigated are the upper and lower jaw [2-5, 9-10]. The focus areas that had been studied are specific area around upper and lower jaw such as around palatal, distal and mesial [2-3]. Besides that area around the teeth (molar and biscuspid) [4-5] and tooth supporting structure such as periodotal ligament space (PDL) [2][5] and lamina dura [3-4] also are the main interest of the investigations. These researches correlates the abnormality pathologies in Ozen [11] which are periapical radiolucency, widen periodontal ligament space and loss of lamina dura. These pathologies are the symptom for the existence of periapical disease [11].

Most surveys include four to five dentists [3], radiologist [2][4] and post graduate with experiences in oral and maxillofacial radiology including digital radiography [5]. The number of image samples used ranges between 12 - 42 of panoramic radiographs [2], periapical digital radiographs [3] [5], interproximal radiographs [4] and bitewing [5] images. Overall results of these works support the idea that digitally enhanced images do provide extra information for dentists [2-3]. The twin-view reading experiments show that it helps improve quality of periapical diseases examinations [4-5]. However these studies compared the overall quality of the non-process images and enhanced images but none had based their assessment on the ability of detecting abnormalities.

Therefore this paper proposes to explore the diagnostic potential between the original and enhanced image by compound enhancement algorithms (CEA), namely

SAHE and SCLAHE. The CEA is the combination of sharpening function (type of high pass filter) and contrast enhancement.

3 Material and Method

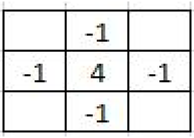
3.1 Material

Thirty intra-oral periapical radiographs are obtained using Planmeca Intra Oral machine from Faculty of Dentistry UiTM Shah Alam. The questionnaire is designed by aligning the original image, the image enhanced with SAHE and SCLAHE in a row (termed as twin-view [3-4]). The images are rated using Riker scale. The subject of the research includes three dentists with experiences ranging between six to fifteen years. This study already received ethical approval by University Technology MARA Ethical Committee (reference No: 600-RMI (5/1/6).

3.2 Method

The methodology consists of three phases; image processing phase; survey phase and finally statistical measurements phase.

The first phase involved image processing processes. SCLAHE consists of two steps; sharpening filter and CLAHE enhancement. Sharpening algorithm is used to sharpen the outline of the periapical features [13] and enhanced bone structure [2]. Laplacian filter is used to perform image sharpening process. It detects the outlines of the objects by convolving a mask with a matrix centered on a target pixel. The Laplacian detects the edge using a mask as in Fig. 1 [13].



	-1	
-1	4	-1
	-1	

Fig. 1. Laplacian Edge Detection Mask

CLAHE on the other hand reduces noise that arises from adaptive histogram equalization (AHE). This technique eliminates the random noise introduced during the AHE process by limiting the maximum slope of the grey scale transform function. The slope of the cumulative distribution function is determined by the bin counts. Large bin count will result in more slopes. Thresholding (clipping) the maximum histogram count, can limit the number of slopes [14].

The second phase involves a survey of dentists' perception ratings on original image and image enhanced with SAHE and SCLAHE. In this phase, the dentist had to classify the presence of periapical radiolucency, the presence of widen periodontal ligament space (widen PDLs) and the presence of loss of lamina dura in the dental images based on the specification in Table 1, Table 2 and Table 3.

Table 1. Rating criteria for detection of the presence of periapical radiolucency

Class	Description
1	Periapical radiolucency detected
2	No periapical radiolucency detected but other abnormality detected
3	No periapical radiolucency detected and no abnormality detected

Table 2. Rating criteria for detection of the presence of widen periodontal ligamen space

Class	Description
1	Widen periodontal ligament space detected
2	No widen periodontal ligament space detected but other abnormality detected
3	No widen periodontal ligament space detected and no abnormality detected

Table 3. Rating tcriteria for detection of the presence of loss of lamina dura

Class	Description
1	Loss of lamina dura detected
2	No loss of lamina dura detected but other abnormality detected
3	No loss of lamina dura detected and no abnormality detected

Explanation of each of the class is as follow; Class 1 is for the pathology that is clearly detected. Class 2 refer to no specified pathology appear in the image but other abnormality detected. Finally class 3 refers to none of the particular pathology as well as other pathologies are detected. Class 3 possibly will be a sign that the teeth were either healthy (since no lesion could be observed) or the image quality is not good at all since it cannot show any lesion clearly.

Finally in the last phase the changes in the image appearance are measured statistically. CII, SNR and RMSE are used to measure the quantitative values between SAHE and SCLAHE.

CII is the popular index used by radiologist to check visibility of lesions in radiographs [15]. It is calculated by $C_{processes}/C_{original}$ where both are the contrast values for the region of interest in the enhanced images and original images respectively [15]. C is defined as in the following equation;

$$C = (f - b)/(f + b) \tag{1}$$

where f is the mean gray-level value of the image and b is the mean gray-level value of the background [16].

SNR is the measurement of the signal out of noise in an image and is calculated by the ratio of the signal standard deviation to the noise standard deviation. The SNR equation is as follows;

$$SNR = 10.\log_{10}\left[\frac{\sum_{0}^{n_x-1}\sum_{0}^{n_y-1}\left[r(x,y)\right]^2}{\sum_{0}^{n_x-1}\sum_{0}^{n_y-1}\left[r(x,y)-t(x,y)\right]^2}\right] \tag{2}$$

The problem in quantifying and improving the enhancement method is that one must able to separate noise from signal. That's how the importance of SNR is. It provides measurement for image quality in term of image details and checks the performance of the enhancement methods [17].

Finally the average magnitude of error in the enhanced image based on the original image is calculated by RMSE and the lower value is the better. [18]. RMSE equation is as follows;

$$RMSE = \sqrt{\frac{1}{n_x n_y}.\sum_{0}^{n_x-1}\sum_{0}^{n_y-1}\left[r(x,y)-t(x,y)\right]^2} \tag{3}$$

4 Result

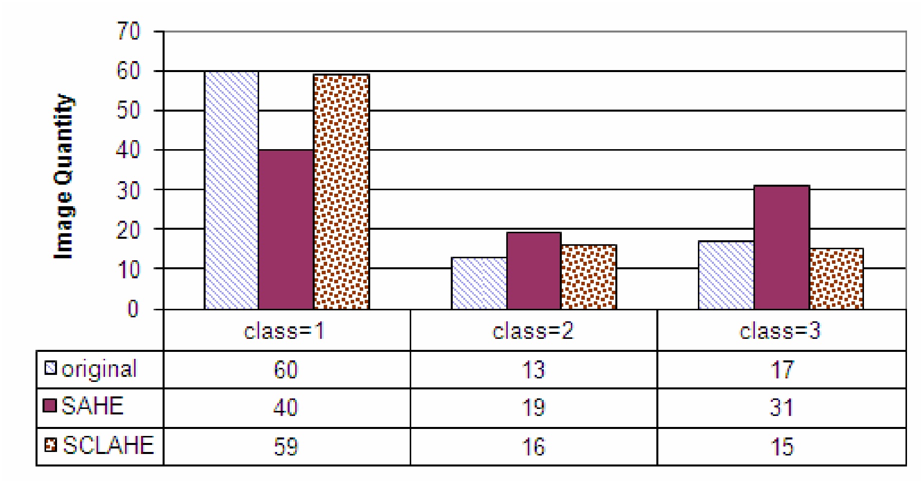
Results are discussed based on dentists' perceptions and statistical analysis of CII, SNR and RMSE of widen periodontal ligament space only. Fig. 2 - 4 show the results of dentists' perception towards the three pathologies; periapical radiolucency, widen periodontal ligament space (PDLs) and loss of lamina dura. Fig. 5 is the clearly detected pathologies or class = 1 only results. The results of Fig. 2 - 5 are based on thirty images. These results are based on each image that get score 1, 2 or 3 based on each dentists' perception. Table IV is the result of widen periodontal ligament space statistical value of each image compared to the dentists' evaluations. This result is based on only twenty six images (due to some technical errors, only twenty six images are able to be extracted for statistical values). Due to limitation of space, the other two abnormality statistical values cannot be reported here.

Fig. 2 shows that for periapical radiolucency, out of 90 observations, there are clearly observed (class = 1) periapical radiolucency abnormality as follows; 60 images by original images, 40 by imaged enhanced by SAHE and 59 by image enhanced by SCLAHE. As for class equal 2 and 3, SAHE show largest values compared to others.

Fig. 3 shows the population of widen PDLs in the sample. It shows that out of 90 observations, for clearly observed widen PDLs (class = 1), SCLAHE got highest values (67), compared to original (62) and SAHE (53). For class 2, original are at par with SAHE (11) and SCLAHE score only 8. As for class 3, SAHE is the highest.

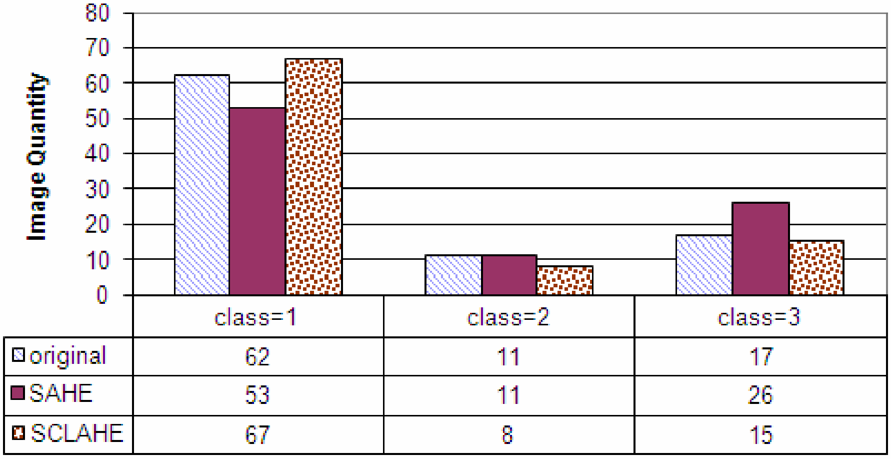
Fig. 4 shows the population of loss of lamina dura. The table indicated that SCLAHE almost at par (61) with the original image (62), while SAHE is the lowest (47).

Fig.5 records only the clearly detected pathology (class = 1). It shows that for periapical radiolucency, original score the best (60) however SCLAHE score only one point behind (59) and SAHE score the least (40). As for widen PDLs, SCLAHE (67) scores more than original image (62) and SAHE (53). Finally for loss of lamina dura, original scores the best (62) but SCLAHE only one point behind (59) and SAHE score only 47.



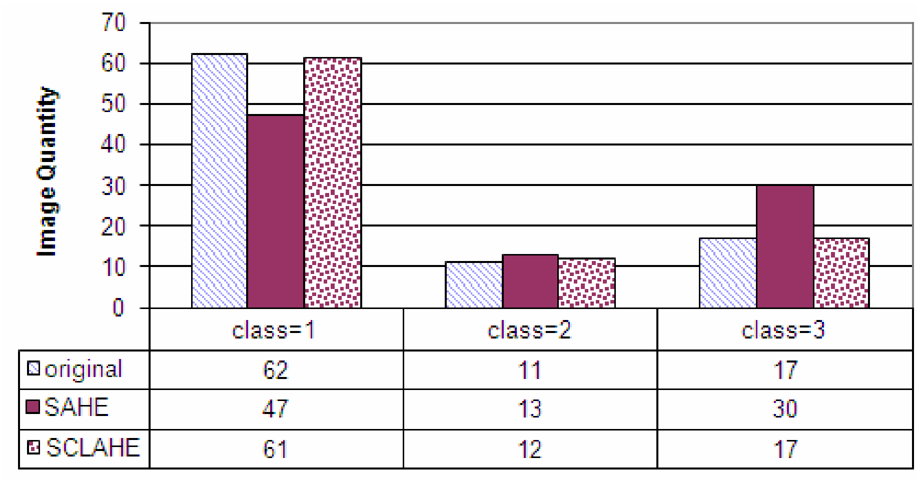
	class=1	class=2	class=3
▨ original	60	13	17
■ SAHE	40	19	31
▩ SCLAHE	59	16	15

Fig. 2. Periapical Radiolucency Abnormality ranking



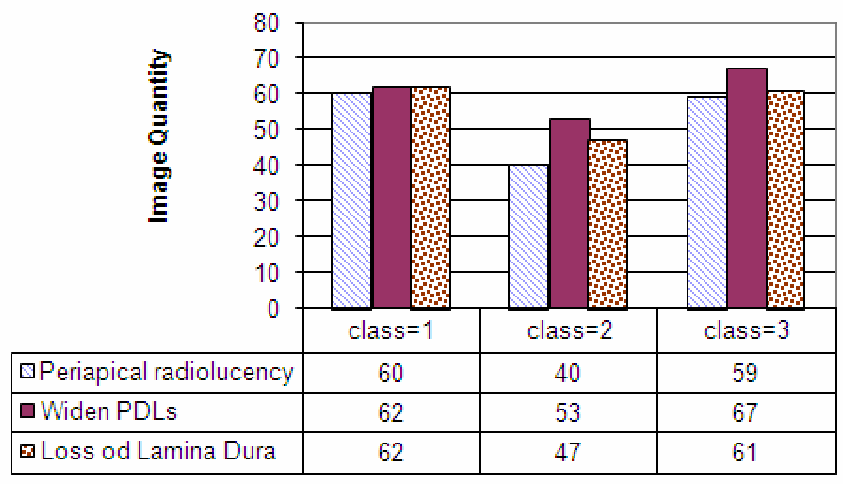
	class=1	class=2	class=3
▨ original	62	11	17
■ SAHE	53	11	26
▩ SCLAHE	67	8	15

Fig. 3. Widen PDLs Abnormality ranking



	class=1	class=2	class=3
▨ original	62	11	17
■ SAHE	47	13	30
⊡ SCLAHE	61	12	17

Fig. 4. Loss of Lamina Dura Abnormality ranking



	class=1	class=2	class=3
▨ Periapical radiolucency	60	40	59
■ Widen PDLs	62	53	67
⊡ Loss od Lamina Dura	62	47	61

Fig. 5. Dentists' perception on clearly detected abnormalities

Due to the limitation of space, the results only limited to widen PDLs. Table IV are the results of statistical values of the widen PDLs pathology for each images (ImgID) compared to the dentists' evaluations. Mode column indicates the highest value rated by the dentists' for original images, SAHE and SCLAHE. The 1, 2 and 3 is the class rating by dentists' for each image as in Table 1, Table 2 and Table 3. The sorting of these tables are based on original mode column. For the row that has the value #N/A, shows that all the dentists have different rating. Statistical values are only for SAHE and SCLAHE as the calculations are based on original images.

The observation of the mode values are as follows; for original images, for score equal to one, original get 20, SCLAHE get 21 and SAHE get 17. For score equal to two, SAHE get three and original and SCLAHE get the same (1). For score equal 3, SAHE get the highest (6), original get 4 and SCLAHE get 3. Finally for #NA, SAHE get none and original and SCLAHE get the same (1). Based on mode values, it shows that only ImgID 11 shows change in rating from 3(original image) to 1 (SAHE and SCLAHE).

The statistical values are calculated based on original images hence the comparison is only between SAHE and SCLAHE performance. The observation of mode that change from rating 3 in SAHE to 1 in SCLAHE are as follows; ImgID 3 and ImgID 6. As for the change in rating from 2 in SAHE to 1 in SCLAHE are as follows; ImgID 8 and ImgID 20. Others values are the same between the two. By comparing the CII and RMSE values between the two CEAs it shows the trend that SAHE has higher values than SCLAHE. As for SNR, SCALHE values are higher than SAHE.

Table 4. The mode of dentists‘ rating and statistical value for the presence of widen Periodontal ligamanet space

ImgID	*Mode Original*	*Mode SAHE*	*CII SAHE*	*SNR SAHE*	*RMSE SAHE*	*Mode SCLAHE*	*CII SCLAHE*	*SNR SCLAHE*	*RMSE SCLAHE*
1	1	1	5.05	7.70	111.18	1	2.88	12.67	67.63
2	1	1	2.98	9.60	91.91	1	1.12	23.02	24.00
3	1	3	15.83	6.25	128.49	1	1.74	22.38	25.59
4	1	1	4.74	9.60	97.65	1	1.21	22.17	27.77
5	1	1	3.47	8.79	99.66	1	1.21	22.18	26.13
6	1	3	22.40	5.58	137.31	1	1.97	22.44	25.44
7	1	1	3.47	9.69	96.74	1	1.23	22.23	27.62
8	1	2	6.72	8.49	109.05	1	1.51	20.33	33.40
9	1	1	2.54	10.16	92.33	1	1.15	21.87	28.62
12	1	1	7.62	11.75	78.75	1	1.39	17.14	45.96
13	1	1	2.90	12.93	69.73	1	1.36	20.27	33.46
15	1	1	1.70	14.05	62.31	1	1.23	18.42	40.27
17	1	1	8.41	10.29	90.79	1	1.49	16.22	50.17
18	1	1	9.11	11.11	83.63	1	1.41	17.01	46.34
19	1	1	15.34	10.63	87.75	1	1.37	17.44	44.42
20	1	2	18.34	8.54	108.11	1	1.85	14.71	58.34
21	1	1	9.81	10.15	92.01	1	1.82	15.13	55.96
22	1	1	6.17	10.56	88.32	1	1.54	16.59	48.34
23	1	1	12.30	9.16	101.58	1	1.68	16.10	50.77
26	1	1	12.52	6.42	126.26	1	1.89	20.48	30.96
10	2	2	0.95	15.36	54.69	2	0.90	20.59	32.42
11	3	1	8.27	10.56	88.36	1	1.45	15.73	52.67
14	3	3	3.16	11.38	81.38	3	1.55	15.66	53.06
24	3	3	1.03	15.26	55.20	3	0.97	20.03	34.27
25	3	3	1.47	14.13	61.85	3	1.13	20.27	33.47
16	#N/A	3	6.05	11.51	80.37	#N/A	1.47	17.29	45.08

5 Discussion

The present study evaluated the performance of compound enhancement algorithm (CEA) namely SAHE and SCLAHE compared to original intra-oral dental radiographs focusing on two aims; 1) dentists' subjective visualization evaluation on the existence of pathologies in pre and post processing and 2) Correlation between dentists' perception and statistical measurement for each image of widen PDLs abnormality.

Figure 2 – 4 shows that dentists able to clearly observed more periapical radiolucency and loss of lamina dura on original images compares to the CEA. As for widen PDLs, SCLAHE produces clearer visual compares to original and SAHE. However performance of SCLAHE is actually quite good as original image as the difference between original and SCLAHE is only 1 (refer to Class = 1, in Fig. 2 and Fig.4). On the other hand, SAHE performance is higher in Class=2, which indicated maybe SAHE is good in detecting other or different pathologies other than periapical radiolucency, widen PDLs and loss of lamina dura.

Figure 5 indicates clearly the performance of CEA over original images in clearly detected pathologies. It shows that dentists prefer original images other than enhanced image. SCLAHE however overcome original images in detecting widen PDLs. This result is in line with the general conclusion of many studies reporting that image processing techniques do improved diagnostic accuracy and may assist dentists in deciding the most suitable treatment for patients [2-5][9-10]. Sharpening function that been applied before CLAHE also plays an important role in enhancing the edges and able to enhance bone structures [2]. Enhanced edges are often more pleasing to human visual system than original images [2][8]. The SCLAHE performance had been reported in [19-21] as better than original image in clearly detecting widen periodontal ligament space [19-21] and loss of lamina dura[19][21]. However the detection of periapical radiolucency of SCLAHE is as par as original images [19-21]. Still none of the papers did any quantitative measurement of the evaluation between the non-processed/original and processes images. Another issue to be considered regarding this finding is that, the methodology of twin-view might effects the dentists' evaluation since they are used to original images thus influence their decisions.

Table IV is the values of mode for dentists' evaluations for widen PDLs abnormality compared to statistical values of CII, SNR and RMSE. These measurements are based on original images as reference. Focusing on ImgID 3 and ImgID 6 higher CII values does not given the dentist' better visual as they chose SCLAHE is more clear (mode =1) than SAHE (mode=3) in detecting the widen PDLs abnormality. As for SNR, the finding is in line with the theory that better signal than noise is in SCLAHE than in SAHE [17]. Looking at RMSE values, the higher values for SAHE shows that the magnitude of errors in this CEA is more than in SCLAHE thus producing lower SNR values [18].

Therefore since the overall results show that SCLAHE values of SNR is higher, but the RMSE is lower, and consistence with the dentists' evaluation mode value, it can be concluded that SCLAHE is superior than SAHE in providing better visualization of the intra-oral dental radiograph.

6 Conclusion and Future Works

In conclusion this work shows that image processing techniques are able to enhance the image subjective quality and providing better information for dentists. In comparison between the performance of original images and CEA, it shows that dentists still prefer original images in detecting periapical radiolucency and loss of lamina dura. However, since the method used is twin-view, bias might effects dentists' evaluation, since they are applied based on the original images. However SCLAHE is almost at par with original images in detecting both pathologies. As for detecting widen periodontal ligament space, SCLAHE able to overcome original images as well as SAHE.

The future work aims to restructuring for a new questionnaire that avoid bias as well as getting more respondents for better and more conclusive results.

Acknowledgment

The authors acknowledge with gratitude the great support from the management of Faculty of Dentistry, UiTM Shah Alam and Universiti Sains Malaysia, Kubang Kerian. Dr Zurairah Berahim and Dr Kasmawati @ Norhidayati Mokhtar that spend their time to answer the questionnaires. And also to Mr Mohd Aziz Farhan Rosli, a radiographer in the department of Dentistry, UiTM Shah Alam. Without their generosity, the collection of periapical dental radiograph images and answering the questionnaires would not be possible. This research received E-science fund under the Ministry of Science Technology and Innovation (MOSTI) Malaysia.

References

1. Stheeman, S.E., Mileman, P.A., Van Hof, M.A., Van Der Stelt, P.F.: Diagnosis confidence and accuracy of treatment decisions for radiopaque periapical lesions. International Endodontic Journal 28, 121–123 (1995)
2. Baksi, B., Alpz, E., Sogur, E., Mert, A.: Perception of anatomical structures in digitally filtered and conventional panoramic radiographs: a clinical evaluation. Dentomaxillofacial Radiology 39, 424–430 (2010)
3. Mehdizadeh, M., Dolatyar, S.: Study of Effect of Adaptive Histogram Equalization on Image Quality in Digital Preapical Image in Pre Apex Area. Research Journal of Biological Science 4(8), 922–924 (2009)
4. Alves, W.E.G.W., Ono, E., Tanaka, J.L.O., Filho, E.M., Moraes, L.C., Moraes, M.E.L., Castilho, J.C.M.: Influence of image filters on the reproducibility of measurements of aveolar bone loss. Journal of Applied Oral Science 4(6), 415–420 (2006)
5. Sund, T., Moystad, A.: Sliding window adaptive histogram eqaulization in intraoral radiographs: effect on image quality. Dentomaxillofacial Radiology 35, 133–138 (2006)
6. Regezi, J.A.: Periapical Disease: Spectrum and Differentiating Features. Journal of The California Dental Association (1999)
7. Parks, E.T.: A Guide to Digital Radiographic Diagnosis: From Panoramic, to Periapicals, to Cone Beam CT',
   http://www.dentalcompare.com/featuredarticle.asp?articleid=169

8. Zhou, J.D., Abdel-Mottaleb, M.: A content-based system for human identification based on bitewing dental x-ray images. Pattern Recognition 38, 2132–2142 (2005)
9. Yalcinkaya, S., Kunze, A., Willian, R., Thoms, M., Becker, J.: Subjective image quality of digitally filtered radiograph acquired by the Durr Vistascan system compared with conventional radiographs. Oral Surg. Oral Med. Pathol. Oral Radiol. Endod. 101, 643–651 (2006)
10. Gijbels, F., Meyer, A.M.D., Serhal, C.B., Bossche, C.V.d., Declerck, J., Persoons, M., Jacob, R.: The subjective image quality of direct digital and conventional panoramic radiography. Journal of Clinical Oral Investigation 4, 162–167 (2000)
11. Ozen, T., Cebeci, A., Paksoy, C.S.: Interpretation of chemically created periapical lesion using 2 different dental cone-beam computerized tomography units, and intraoral digital sensor, and conventional film. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. 107(3), 426–432 (2009)
12. Whaites, E.: Essentials of Dental Radiographic and Radiology. Elsevier, Amsterdam (2003)
13. Allen, B., Wilkinson, M.: Parallel Programming, Techniques and Applications Using Networked Workstations and Parallel Computers. Pearson, London (2005)
14. Poulist, J., Aubin, M.: Contrast Limited Adaptive Histogram Equalization(CLAHE) , http://radonc.ucsf.edu/research_group/jpouliot/Tutorial/HU/Lesson7.htm (accessed on June 4, 2010)
15. Yoon, J.H., Ro, Y.M.: Enhancement of the Contrast in Mammographic Images Using the Homomorphic Filter Method. IEICE Trans. Inf. & Syst. E85-D, 289–303 (2002)
16. Bankman, I.N.: Handbook of Medical Image Processing and Analysis. Academic Press, London (2008)
17. Sijbers, J., Scheunders, P., Bonnet, N., Duck, D.V., Raman, E.: Quantification and Improvement of the Signal-to-Noise Ratio in a Magnetic Resonance Image Acquisition Procedure. Magnetic Resonance Imaging 14, 1157–1163 (1996)
18. Thangavel, K., Manavalan, R., Aroquiaraj, I.L.: Removal of Speckle Noise from Ultrasound Medical Image based on Special Filters: Comparative Study. ICGST-GVIP Journal 9(III), 25–32 (2009)
19. Ahmad, S.A., Taib, M.N., Khalid, N.E., Ahmad, R., Taib, H.: The Effect of Sharp Contrast-Limited Adaptive Histogram Equalization (SCLAHE) on Intra-oral Dental Radiograph Images. In: 2010 IEEE EMBS Conference on Biomedical Engineering & Sciences, IECBES (2010)
20. Ahmad, S.A., Taib, M.N., Khalid, N.E., Ahmad, R., Taib, H.: A comparison of Image Enhancement Techniques for Dental X-ray Image Interpretation. In: 2010 International Conference on Computer and Computational Intelligence (ICCCI), vol. 1, pp. v2-141–v2-145 (2010)
21. Ahmad, S.A., Taib, M.N., Khalid, N.E., Ahmad, R., Taib, H.: Performance of Compound Enhancement Algorithms on Dental Radiograph Images. In: International Conference on Medical Image and Signal Computing ICMISC 2011, Penang, Malaysia, February 23-25 (2011)

Performance Study of Two-Dimensional Orthogonal Systolic Array

Ahmad Husni Mohd Shapri, Norazeani Abdul Rahman,
and Mohamad Halim Abd. Wahid

School of Microelectronic Engineering, Universiti Malaysia Perlis,
Blok A, Kompleks Pusat Pengajian Jejawi 1,
02600 Jejawi, Perlis, Malaysia
{ahmadhusni,azeani,mhalim}@unimap.edu.my

Abstract. The systolic array implementation of artificial neural networks is one of the ideal solutions to communication problems generated by highly interconnected neurons. A systolic array is an arrangement of processors in an array where data flows synchronously across the array between neighbours, usually with different data flowing in different directions. The simulation of systolic array for matrix multiplication is the practical application in order to evaluate the performance of systolic array. In this paper, a two-dimensional orthogonal systolic array for matrix multiplication is presented. Perl scripting language is used to simulate a two-dimensional orthogonal systolic array compared to conventional matrix multiplication in terms of average execution time. The comparison is made using matrices of size 5xM versus Mx5 which M ranges from 1 to 10, 10 to 100 and 100 to 1000. The orthogonal systolic array results show better average execution time when M is more than 30 compared to conventional matrix multiplication when the size of the matrix multiplication is increased.

Keywords: orthogonal systolic array; matrix multiplication; perl scripting.

1 Introduction

Matrix multiplication is the operation of multiplying a matrix with either a scalar or another matrix. The ordinary matrix product is most often used and the most important way to multiply matrices. It is defined between two matrices only if the width of the first matrix equals the height of the second matrix. Multiplying the *m x n* matrix with *n x p* matrix will result in *m x p* matrix. If many matrices are multiplied together and their dimensions are written in a list in order, e.g. *m x n, n x p, p x q* and *q x r,* the size of the result is given by the first and the last numbers which is *m x r.* The values surrounding each comma must match for the result to be defined.

A number of systolic algorithms are available for matrix multiplication, the basic computation involved in the operation of a neural network. Using these, many systolic algorithms have been formulated for the implementation of neural networks [1]. A systolic array is composed of matrix-like rows of data processing units called cells or processing elements. Each cell shares the information with its neighbours immediately after processing. The systolic array is often rectangular where data flows across

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 567–574, 2011.

the array between neighbour cells, often with different data flowing in different directions. Kung *et. al.* have proposed a unified systolic architecture for the implementation of neural network models [2]. It has been shown that the proper ordering of the elements of the weight matrix makes it possible to design a cascaded dependency graph for consecutive matrix multiplication, which requires the directions of data movement at both the input and the output of the dependency graph to be identical. Using this cascaded dependency graph, the computations in both the recall and the learning iterations of a back-propagation algorithm have been mapped onto a ring systolic array.

A similar mapping strategy has been used in [3] for mapping the recursive back-propagation network and the hidden Markov model onto the ring systolic array. The main disadvantage of the above implementations is the presence of spiral communication links. In [2], a two-dimensional array is used to map the synaptic weights of individual weight layers in the neural network. By placing the arrays corresponding to adjacent weight layers side by side, both the recall and the learning phases of the back-propagation algorithm can be executed efficiently. But, as the directions of data movement at the output and the input of each array are different, this leads to a very non-uniform design.

2 Background

There are also other applications of the systolic array multiplication which are matrix polynomial and powers of a matrix [4]. A matrix polynomial is a matrix whose elements are univariate or multivariate polynomials. A slight modification of the design allow the creation of matrix multiply-and-add step of the form $X^{(n)} \leftarrow X^{(n-1)}A+B$. The accumulation of B can be done on the fly when the values $X^{(n-1)}A$ reach the left row of the array, as depicted by Fig. 1. Consider the calculation of the matrix polynomial $P=\sum_{k=0}B_kA^k$ where B_k and A are n x n matrices. The algorithm is also valid when the matrices are n x p, or when the coefficients B_k are vectors. Using Horner's rule, P can be computed by the following iterative scheme:

$$X^{(0)} \quad = B_N \tag{1}$$

$$X^{(k)} \quad = X^{(k-1)}A+B_{N-k} \tag{2}$$

$$P \quad \equiv X^{(N)} \tag{3}$$

Therefore, P can be computed in exactly $2n(N+1)-1$ steps on the systolic array of Fig. 1. A well-known efficient algorithm for the computation of matrix powers $P = A^N$ is described by Knuth [5] and Lamagna [6]. It consists in using the binary representation of N to control a sequence or multiply by A steps. Although this method is not optimal, it has the greatest advantage, compared to others, that is does not require an important storage. This property makes it well-suited for a systolic realization.

More precisely, let $N^P\ N^{p-}\ldots N^0$ be the binary representation of N, and let C_q, C_{q-1},...C_0 be the new sequence obtained by rewriting each N_k as SX if N_k =1 or as S if N_k =1 or as S if $N_k = 0$, and by discarding the first pair of letters SX. C is used to control a sequence of operations and is interpreted from the left to right, each S meaning square the current matrix and each X equal to multiply the current matrix by A. For example, decimal 19 is equal to binary 10011 will be rewritten as $SSSXSX$ and the sequence of computations will be A^2, A^4, A^8, A^9, A^{18}, A^{19}. Then the following iteration scheme

$$X^{(0)} = A \tag{4}$$

$$X^{(n)} = (X^{(n-1)})^2 \; if \; C_{q-r} = S \; else \; X^{(n-1)} A \tag{5}$$

$$P \quad \equiv X^{(q)} \tag{6}$$

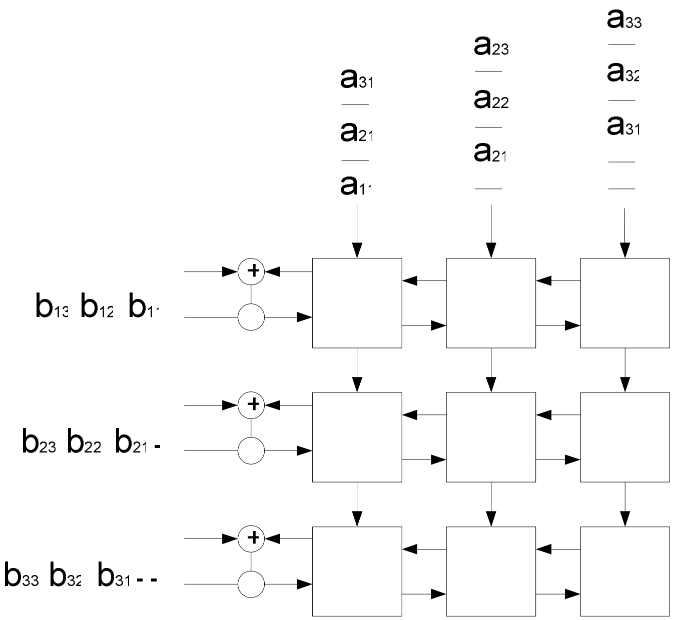


Fig. 1. Systolic array for the matrix polynomial calculation

The basic calculation to be done is either a square or a matrix multiplication. Again, a simple modification of our design enables us to square a matrix. The basic idea is to route the result to the upper row of the array too. Depending on the value of C_{q-1}, A or $X^{(n-1)}$ will be fed into the array from the top to bottom. Elementary movements along the vector (1, 0, -1) is compatible with the timing-function $t(i, j, k)=i+j+k-3$ and the resulting design is shown by Fig. 2. As the binary representation of N has $[log_2N]+1$ bits, q is majored by $2[log_2+N]$. Hence, the calculation of A^N takes a maximum of $2n(2Log_2N+1)-1$ steps on a n^2 mesh connected array.

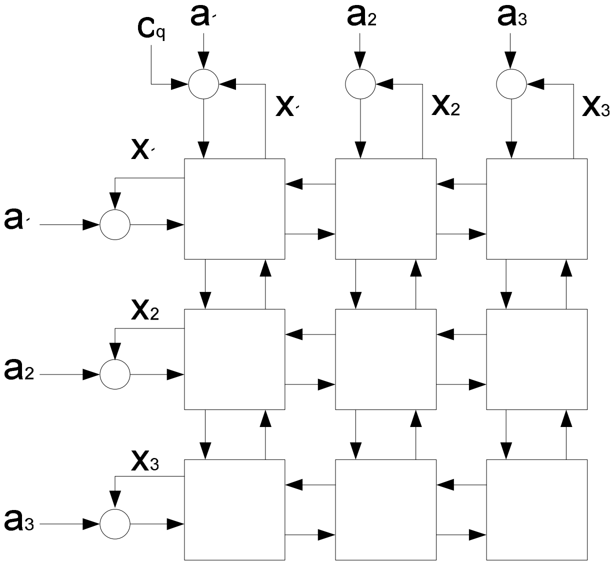


Fig. 2. Systolic array for the powers of matrix calculation.

3 Methodology

We have created a simulation program using perl scripting language that simulates the orthogonal systolic array and conventional matrix multiplication. The goal of the simulator design is to develop an accurate, fast and stable simulator based on an orthogonal systolic array and conventional matrix multiplication.

3.1 Perl Scripting Language

Scripting languages such as perl, tcl, tk and python are the main achievement of the open source movement. Perl is a popular and widely-used cross-platform programming language. In particular, its flexibility can be fully utilized to become a powerful simulator. There are no rules about indentation, newlines, etc. Most lines end with semicolons, but not everything has to. Most things do not have to be declared, except for a couple of things that do [7]. Perl only has three basic data types: scalars, arrays, and hashes. It stores numbers internally as either signed integers or double-precision, floating-point values.

Perl is an open source interpreted programming language. It is a control program that understands the semantics of the language and its components, the interpreter executes program components individually as they are encountered in the control flow.

Today this is usually done by first translating the source code into an intermediate representation called bytecode and then interpreting the bytecode. Interpreted execution makes perl flexible, convenient and fast for programming, with some penalty paid in execution speed.

One of the major advantages of perl scripting language is the support for regular expressions. Regular expressions are the key to powerful, flexible, and efficient text processing. Regular expressions are used in several ways in perl. They are used in conditionals to determine whether a string matches a particular pattern. They are also used to find patterns in strings and replace the match with something else. Regular expressions themselves with a general pattern notation allow the designer to describe and parse text [8].

3.2 Two-Dimensional Orthogonal Systolic Array

Systolic array designs have two main characteristics: a) they are flow-simple, where each processor element is used once every clock-cycle and they are locally connected and b) where each processing element is connected to the nearest neighbour. Matrix-multiplication systolic array is well known as the matrix multiplication, $C = AB$ as shown in Fig. 3 can be implemented on two dimensional arrays using orthogonal arrangement in Fig. 4. Numerous designs can be used depending on whether one wants C, A or B to move. Here we consider the simple design where C stays in place. Coefficient c_{ij}, where $1 \leq i, j \geq n$ thus is calculated by cell i, j of a mesh connected array of multiply-and-add processors, as depicted by Fig. 4. Cell i, j contains a register c which is first reset to zero and then accumulates successively products $a_{ik}\, b_{kj}$ where k varies from 1 to n. However this implementation suffers from the drawback that the results do not move and consequently a systolic output scheme has to be designed in order to recover them.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

Fig. 3. Matrix multiplication $C = AB$

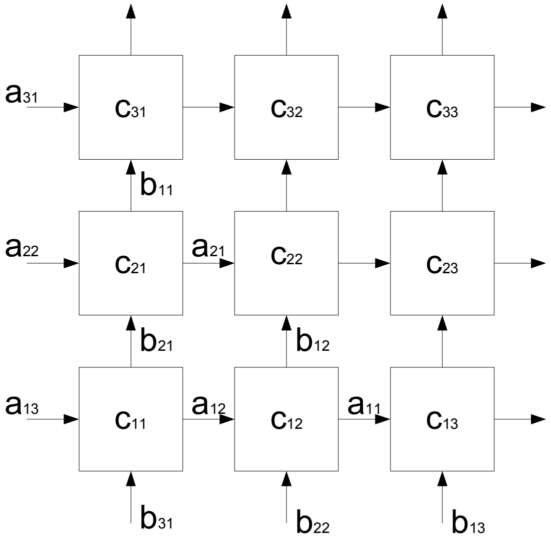


Fig. 4. Orthogonal systolic array for the multiplication of matrices

4 Implementation Details

The simulation program consists of two main scripts which are systolic.pl and run.pl. The systolic.pl is the core of the simulation program. It has five parts, which are listed in Table 1. The run.pl has been developed in order to get the average execution time. The purpose of this script is to execute the systolic.pl in batch mode, which is to get 10 values of execution time for each matrix multiplication. These values can be used to calculate the average execution time. All scripts are executed in Windows® environment by using Cygwin™ terminal.

Table 1. Five parts of systolic.pl

No	Part	Functions
1	random matrix generator	to generate the value for matrices or select the files for matrix inputs.
2	systolic array	to apply systolic array matrix multiplication using orthogonal arrangement.
3	conventional	to apply matrix multiplication using conventional approach.
4	execution time	to capture the execution time for each approach.
5	result	to display the result on screen and record it to files.

The maximum size of output matrix is 5x5. Hence, the matrix size for the simulation is 5xM versus Mx5, where the M value can be divided into three parts which is from 1 to 10, 10 to 100 and 100 to 1000 and the number of processing elements used in the simulation is 25.

5 Result

Fig. 5 shows the average execution times for a matrix multiplication using orthogonal systolic array and conventional approach for matrix values, M, from 1 to 10. At M=1, a big differences of average execution time occurs between these two approaches by

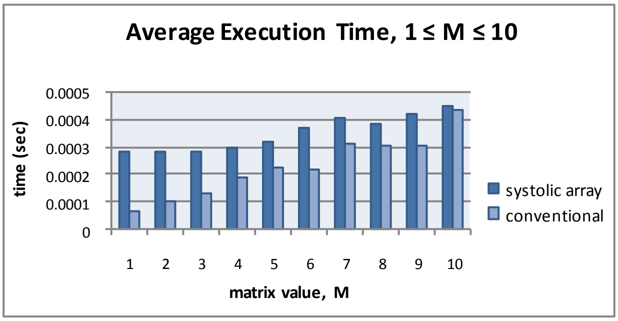


Fig. 5. Average execution time for 1 ≤ M ≤ 10.

61%. While at M=10, only 1.5% difference is seen between the two approaches. From here, we noticed that the orthogonal systolic array needs more time to compute the matrix multiplication for a small matrix size compare to the conventional approach because the matrix was been multiplied using reusable processing elements, which introduced delays.

For matrix size with value M from 10 to 100, as shown in Fig. 6, the average execution times for matrix multiplication using conventional approach is increased. For M=10, the average execution time required for orthogonal systolic array to complete the matrix multiplication is 11.9% higher than conventional approach. When M=20, the differences became smaller and slightly equal to conventional approach. After M=30 and above, the average execution time for conventional approach is higher than the orthogonal systolic array approach.

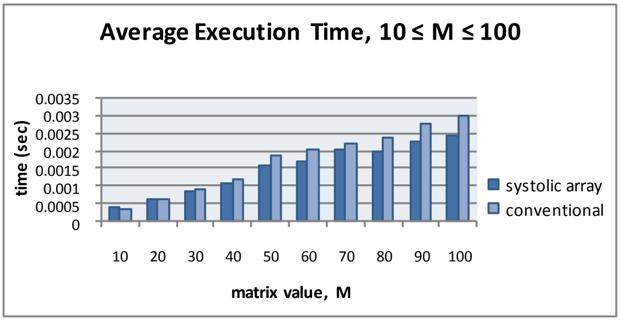


Fig. 6. Average execution time for 10 ≤ M ≤ 100.

In Fig. 7, the average execution time for matrix multiplication using orthogonal systolic array is lower than the conventional approach. The total average execution time for orthogonal systolic array is 150 ms while that for the conventional approach is 174 ms. The percentage increase is up to 7.4% for 100 ≤ M ≤ 1000. The orthogonal systolic array approach is much better than conventional approach for large size matrix multiplications.

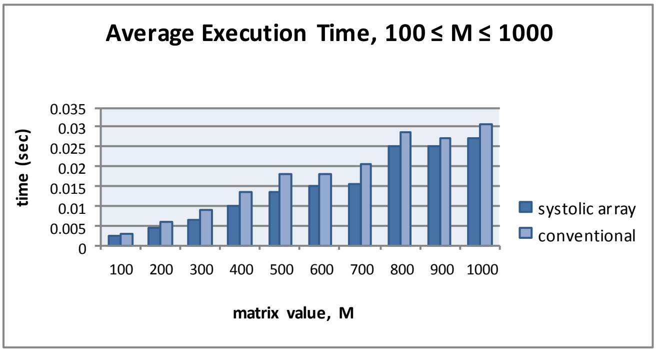


Fig. 7. Average execution time for 100 ≤ M ≤ 1000

6 Conclusion

This paper evaluates the performances of two-dimensional orthogonal systolic array compared to conventional approach for the multiplication of matrices. Three series of evaluation has been completed for M in range of 1 to 10, 10 to 100 and 100 to 1000. From the results, the average execution time was examined and the orthogonal systolic array show a better results compared to the conventional approach. When M=30 and above, the orthogonal systolic array shows less average execution time, which performed better. When M=20, the average execution time between orthogonal systolic array and conventional approach is slightly same and at M=1, major differences of value occurred. The main advantage of the orthogonal systolic array is the processing elements can be reused for a new multiplication without the need of intermediate storage.

In conclusion, an orthogonal systolic array can be used perfectly in order to handle large sizes of matrix multiplication with better performance than conventional approach.

References

1. Sudha, N., Mohan, A.R., Meher, P.K.: Systolic array realization of a neural network-based face recognition system. In: IEEE International Conference on Industrial Electronics and Applications (ICIEA 2008), pp. 1864–1869 (2008)
2. Kung, S.Y., Hwang, J.N.: A unified systolic architecture for artificial neural networks. J. Parallel Distrib. Comput. 6, 358–387 (1989)
3. Hwang, J.N., Vlontzos, J.A., Kung, S.Y.: A systolic neural network architecture for hidden markov models. IEEE Trans. on ASSP 32(12), 1967–1979 (1989)
4. Shapri, A.H.M., Rahman, N.A.: Performance evaluation of systolic array matrix multiplication using scripting language. In: Proc. Regional Conference of Solid State Science and Technology, RCSSST (2009)
5. Knuth, D.: The Art of Computer Programming. Seminumerical Algorithms, vol. 2, pp. 398–422. Addison Wesley, Reading (1969)
6. Lamagna, E.A.: Fast Computer Algebra, vol. 43. IEEE Computer Society Press, Los Alamitos (1982)
7. Spainhour, S., Siever, E., Patwardhan, N.: Perl in a Nutshell, 2nd edn., pp. 43–45. O'Reilly Media, Sebastopol (2002)
8. Friedl, J.E.F.: Mastering Regular Expressions. O'Reilly & Associates, Inc., Sebastopol (1998)
9. Muroga, C.: On a Case of Symbiosis between Systolic Arrays. Integration the VLSI Jurnal 2, 243–253 (1984)
10. Amin, S.A., Evans, D.J.: Systolic array design for low-level image processing. Kybernetes 23 (1994)
11. Chung, J.H., Yoon, H.S., Maeng, S.R.: A Systolic Array Exploiting the Inherent Parallelisms of Artificial Neural Networks. Micro-processing and Microprogramming, vol. 33. Elsevier Science Publishers B. V, Amsterdam (1992)
12. Kane, A.J., Evans, D.J.: An instruction systolic array architecture for neural networks. International Journal of Computer Mathematics 61 (1996)
13. Mahapatraa, S., Mahapatra, R.N.: Mapping of neural network models onto systolic arrays. Journal of Parallel and Distributed Computing 60, 677–689 (2000)

Similarity Approach on Fuzzy Soft Set Based Numerical Data Classification

Bana Handaga and Mustafa Mat Deris

University Tun Hussein Onn Malaysia
handaga.bana@gmail.com, mmustafa@uthm.edu.my

Abstract. Application of soft sets theory for classification of natural textures has been successfully carried out by Mushrif et. al.. However the approach can not be applied in a particular classification problem, such as problem of text classification. In this paper, we propose the new numerical data classification based on similarity fuzzy soft sets. In addition can be applied to text classification, this new fuzzy soft sets classifier (FSSC) can also be used in general numerical data classification. As compare to previous soft sets classifier on seven real data sets experiments, the new proposed approach give high degree of accuracy with low computational complexity.

Keywords: fuzzy soft set theory, numerical data, classification.

1 Introduction

Classification, one of the most popular and significant machine learning areas, is particularly important when a data repository contains samples that can be used as the basis for future decision making: for example, medical diagnosis, credit fraud detection or image detection. Machine learning researchers have already proposed many different types of classification algorithms, including nearest-neighbour methods, decision tree induction, error backpropagation, reinforcement learning, lazy learning, rule-based learning and relatively new addition is statistical learning. From amongst this vast and ever increasing array of classification algorithms, it becomes important to ask the question 'which algorithm should be the first choice for my present classification problem?'

To answer this question, Ali and Smith [2] has conducted research comparing the 8 algorithms/classifiers with 100 different classification problems. The relative weighted performance measures showed that there was no single classifier to solve all 100 classification problems with best performance over the experiments. There have been numerous comparisons of the different classification and prediction methods, and the matter remains a research topic. No single method has been found to be superior over all others for all data sets [8]. This is our motivation to proposed a new classification algorithm, based on soft sets theory.

In 1999, D. Molodtsov [16], introduced the notion of a soft set as a collection of approximate descriptions of an object. This initial description of the object has an approximate nature, and we do not need to introduce the notion of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 575–589, 2011.

exact solution. The absence of any restrictions on the approximate description in soft sets make this theory very convenient and easily applicable in practice. Applications of soft sets in areas ranging from decision problems to texture classification, have surged in recent years [5,12,17,21].

The soft sets theory can work well on the parameters that have a binary number, but still difficult to work with parameters that have a real number. There are many problems in the classification involving real numbers. To overcome this problem, Maji et al. [11] have studied a more general concept, namely the theory of fuzzy soft sets, which can be used with the parameters in the form of real numbers. These results further expand the scope of application soft sets theory. There are two important concepts underlying the application of the theory of soft sets in numerical classification problems. First, the concept of decision-making problems based on the theory of fuzzy soft sets, and the second is the concept of measuring similarity between two fuzzy soft sets.

Based on an application of soft sets in a decision making problem presented by [12], Mushrif et al. [17] presented a novel method for classification of natural textures using the notions of soft set theory, all features on the natural textures consist of a numeric (real) data type, have a value between [0,1] and the algorithm used to classify the natural texture is very similar to the algorithms used by Roy and Maji [21] in the decision-making problems with the theory of fuzzy sets softs. In their experiments, Mushrif et al. used 25 texture classes with 14 texture features. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less in case of soft set method in comparison with Bayes classification method. However, soft sets approach proposed by the [17] can not work properly in particular classification problem, such as problem of text classification. This classifier has a very low accuracy, and even failed to classify text data.

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovering task, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. The study of how to measure the similarity between soft sets have been carried out by Majumdar and Samanta [15] and Kharal [9], then Majumdar and Samanta [14] is also expanding his study to measure the similarity of fuzzy soft set and describe how to used that formula in medical diagnosis to detect wheter an ill person to perform certain is suffering from a disease or not.

This paper proposed a new approach of classification based on fuzzy soft set theory, using concept of similarity between two fuzzy soft sets, we call this classifier as fuzzy soft sets classsifier (FSSC). FSSC, first construct model fuzzy soft sets $(\widetilde{F}, E)$ for each class and construct model fuzzy soft sets $(\widetilde{G}, E)$ for data without class labels. Next find the similarity measure, $S(\widetilde{F}, \widetilde{G})$, between $(\widetilde{F}, E)$ and $(\widetilde{G}, E)$. The new data will be given a class label according to the class with the highest similarity. As compare to (fuzzy) soft sets classification of natural textures by Mushrif et. al. on two real data sets experiments, the new proposed approach give high degree of accuracy with low computational complexity. We use the seven types of data sets from UCI to compare between the proposed fuzzy soft-set classifier to soft set classifier proposed by [17].

The rest of this paper is organized as follows. In the next section, we describes soft set and fuzzy soft sets. In Section 3, describes the concepts of classification base on soft sets theory and we describes our new propose FSSC algorithm in this section. Section 4, describes the classifier evaluation methodology. In Section 5, discuss about experiments to compare between FSSC and soft sets classification algorithm proposed by Mushrif et. al.. In the final section, some concluding comments are presented.

2 Soft Set Theory

In this section, we recall the basic notions of soft sets and fuzzy soft sets. Let U be an initial universe of objects and E_U (simply denoted by E) the set of parameters in relation to objects in U. Parameters are often attributes, characteristics, or properties of objects. Let $P(U)$ denote the power set of U and $A \subseteq E$. Following [16,13], the concept of soft sets is defined as follows.

2.1 Definition of Soft Set

Definition 1. ([16]) Let U be initial universal set and let E be a set of parameters. Let $P(U)$ denote the power set of U. A pair (F, E) is called a soft set over U, if only if F is a mapping given by $F{:}A \rightarrow P(U)$.

By definition, a soft set (F, E) over the universe U can be regarded as a parameterized family of subsets of the universe U, which gives an approximate (soft) description of the objects in U. As pointed in [16], for any parameter E, the subset may be considered as the set of ϵ-approximate elements in the soft set (F, E). It is worth noting that $F(\epsilon)$ may be arbitrary: some of them may be empty, and some may have nonempty intersection [16]. For illustration, Molodtsov considered several examples in [16]. Similar examples were also discussed in [13,1].

From that definition, it is known that fuzzy set introduced by L.A. Zadeh [23] is kind of special soft sets. Let A be a fuzzy set and μ_A be a subject function (μ_A be a mapping of U into [0,1]). To this problem, Maji et al. [11] initiated the study on hybrid structures involving both fuzzy sets and soft sets. They introduced in [11] the notion of fuzzy soft sets, which can be seen as a fuzzy generalization of (crisp) soft sets.

2.2 Definition of Fuzzy Soft Sets

Definition 2. ([11]) Let U be an initial universal set and let E be set of parameters. Let $\widetilde{P}(U)$ denote the power set of all fuzzy subsets of U. Let $A \subset E$. A pair $(\widetilde{F}, E)$ is called a fuzzy soft set over U, where $\widetilde{F}$ is a mapping given by $\widetilde{F}{:}A \rightarrow \widetilde{P}(U)$.

In the above definition, fuzzy subsets in the universe U are used as substitutes for the crisp subsets of U. Hence it is easy to see that every (classical) soft set may be considered as a fuzzy soft set. Generally speaking $\widetilde{F}(\epsilon)$ is a fuzzy subset in U and it is called the fuzzy approximate value set of the parameter ϵ.

It is well known that the notion of fuzzy sets provides a convenient tool for representing vague concepts by allowing partial memberships. In the definition of a fuzzy soft set, fuzzy subsets are used as substitutes for the crisp subsets. Hence every soft set may be considered as a fuzzy soft set. In addition, by analogy with soft sets, one easily sees that every fuzzy soft set can be viewed as an (fuzzy) information system and be represented by a data table with entries belonging to the unit interval [0,1]. For illustration, we consider the following example

Example 3. This example taken from [14], suppose a fuzzy soft set $(\widetilde{F}, E)$ describes attractiveness of the shirts with respect to the given parameters, which the authors are going to wear. $U = \{x_1, x_2, x_3, x_4, x_5\}$ which is the set of all shirts under consideration. Let $\widetilde{P}(U)$ be the collection of all fuzzy subsets of U. Also let E= e_1= "colorful", e_2 = "bright", e_3 = "cheap", e_4 = "warm". Let

$$F(e_1) = \left\{\frac{x_1}{0.5}, \frac{x_2}{0.9}, \frac{x_3}{0}, \frac{x_4}{0}, \frac{x_5}{0}\right\},\ F(e_2) = \left\{\frac{x_1}{1.0}, \frac{x_2}{0.8}, \frac{x_3}{0.7}, \frac{x_4}{0}, \frac{x_5}{0}\right\}$$

$$F(e_3) = \left\{\frac{x_1}{0}, \frac{x_2}{0}, \frac{x_3}{0}, \frac{x_4}{0.6}, \frac{x_5}{0}\right\},\ F(e_4) = \left\{\frac{x_1}{0}, \frac{x_2}{1.0}, \frac{x_3}{0}, \frac{x_4}{0}, \frac{x_5}{0.3}\right\}$$

Then the family $\{\widetilde{F}(e_i), i = 1, 2, 3, 4\}$ of $\widetilde{P}(U)$ is a fuzzy soft set $(\widetilde{F}, E)$. Tabular representation for fuzzy soft sets $(\widetilde{F}, E)$ show in Table 1.

Table 1. Tabular representation of fuzzy Soft-Set $(\widetilde{F}, E)$

$\boldsymbol{U}$	e_1	e_2	e_3	e_4
x_1	0.5	1.0	0	0
x_2	0.9	0.8	0	1.0
x_3	0	0.7	0	0
x_4	0	0	0.6	0
x_5	0.2	0.8	0.8	0.4

There are two important concepts underlying the application of the theory of fuzzy soft sets in numerical classification problems. First, the concept of decision-making problems based on the theory of fuzzy soft sets early proposed by Maji et. al. [21], and the second is the concept of measuring similarity between two fuzzy-soft-sets proposed by Majumdar and Samanta [14]. The following will be explained briefly about these two concepts.

2.3 Fuzzy Soft Set Based Decision Making

We begin this section with a novel algorithm designed for solving fuzzy soft set based decision making problems, which was presented in [21]. We select this algorithm, because this algorithm have similarity with soft sets classifier algorithm proposed by Mushrif et. al. [17] for classification of natural textures. Next Feng [5] show that this algorithm is actually a useful method for selecting the optimal alternative in decision making problems based on fuzzy soft sets, while the counter example given in [10] is not sufficient for concluding that the Roy-Maji method is incorrect.

Algorithm 4. Fuzzy soft sets decision making problem algorithm by Roy-Maji [21].

1. Input the fuzzy soft sets $(\widetilde{F}, A)$, $(\widetilde{G}, B)$ and $(\widetilde{H}, C)$.
2. Input the parameter set P as observed by the observer.
3. Compute the corresponding resultant fuzzy soft set $(\widetilde{S}, P)$ from the fuzzy soft sets $(\widetilde{F}, A)$, $(\widetilde{G}, B)$, $(\widetilde{H}, C)$ and place it in tabular form.
4. Construct the comparison table of the fuzzy soft set $(\widetilde{S}, P)$ and compute r_i and t_i for o_i, $\forall i$.
5. Compute the score s_i of o_i, $\forall i$.
6. The decision is o_k if $s_k = max_i\ s_i$.
7. If k has more than one value then any one of o_k may be chosen.

Roy and Maji [21] pointed out that the object recognition problem may be viewed as a multi-observer decision making problem, where the final identification of the object is based on the set of inputs from different observers who provide the overall object characterization in terms of diverse sets of parameters. The above Algorithm-4 gives solutions to the recognition problem by means of fuzzy soft sets. This method involves construction of comparison table from the resultant fuzzy soft set and the final optimal decision is taken based on the maximum score computed from the comparison table.

The comparison table is a square table in which rows and columns both are labelled by the object names $o_1, o_2, \ldots, o_n$ of the universe, and the entries c_{ij} indicate the number of parameters for which the membership value of o_i exceeds or equals the membership value of o_j. Clearly, $0 \leq c_{ij} \leq m$, where m is the number of parameters. The row-sum r_i of an object o_i is computed by

$$r_i = \sum_{i=1}^{n} c_{ij} \tag{1}$$

Similarly the column-sum t_j of an object o_j is computed by

$$t_j = \sum_{j=1}^{n} c_{ij} \tag{2}$$

Finally, the score s_i of an object o_i is defined by

$$s_i = r_i - t_i \tag{3}$$

The basic idea of Algorithm-4 was illustrated in [21] by a concrete example (see Section 4 of [21] for details).

2.4 Similarity between Two Soft Sets

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovering task, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. Several researchers have studied the problem of similarity measurement between fuzzy sets, fuzzy numbers and vague sets. Recently Majumdar and Samanta [14,15] have studied the similarity measure of soft sets and fuzzy soft sets. Similarity measures have extensive application in several areas such as pattern recognition, image processing, region extraction, coding theory etc.

In General Fuzzy Soft Set (GFSS) [14] explain similarity between the two GFSS as follow. Let $U = \{x_1, x_2, \ldots, x_n\}$ be the universal set of elements and $E = \{e_1, e_2, \ldots, e_m\}$ be the universal set of parameters. Let F_ρ and G_δ be two GFSS over the parametrized universe (U, E). Hence $F_\rho = \{(F(e_i), \rho(e_i)), i = 1, 2, \ldots, m\}$ and $G_\delta = \{(G(e_i), \delta(e_i)), i = 1, 2, \ldots, m\}$.

Thus $\widetilde{F} = \{F(e_i), i = 1, 2, \ldots, m\}$ and $\widetilde{G} = \{G(e_i), i = 1, 2, \ldots, m\}$ are two families of fuzzy soft sets.

Now the similarity between $\widetilde{F}$ and $\widetilde{G}$ is found first and it is denoted by $M(\widetilde{F}, \widetilde{G})$. Next the similarity between the two fuzzy sets ρ and δ is found and is denoted by $m(\rho, \delta)$. Then the similarity between the two GFSS F_ρ and G_δ is denoted as $\mathcal{S}(F_\rho, G_\delta) = M(\widetilde{F}, \widetilde{G}) \cdot m(\rho, \delta)$. Here $M(\widetilde{F}, \widetilde{G}) = \max_i M_i(\widetilde{F}, \widetilde{G})$, where

$$M_i(\widetilde{F}, \widetilde{G}) = 1 - \frac{\sum_{j=1}^{n} \left| \widetilde{F}_{ij} - \widetilde{G}_{ij} \right|}{\sum_{j=1}^{n} \left(\widetilde{F}_{ij} + \widetilde{G}_{ij} \right)},\ \widetilde{F}_{ij} = \mu_{\widetilde{F}(e_i)}(x_j)\ and\ \widetilde{G}_{ij} = \mu_{\widetilde{G}(e_i)}(x_j)$$

Also

$$m(\rho, \delta) = 1 - \frac{\sum |\rho_i - \delta_i|}{\sum (\rho_i + \delta_i)},\ where\ \rho_i = \rho(e_i)\ and\ \delta_i = \delta(e_i).$$

if we used universal fuzzy soft set then $\rho = \delta = 1$ and $m(\rho, \delta) = 1$, now formula for similarity is

$$\mathcal{S}(F_\rho, G_\delta) = M_i(\widetilde{F}, \widetilde{G}) = 1 - \frac{\sum_{j=1}^{n} \left| \widetilde{F}_{ij} - \widetilde{G}_{ij} \right|}{\sum_{j=1}^{n} \left(\widetilde{F}_{ij} + \widetilde{G}_{ij} \right)}, \tag{4}$$

where $\widetilde{F}_{ij} = \mu_{\widetilde{F}(e_i)}(x_j)\ and\ \widetilde{G}_{ij} = \mu_{\widetilde{G}(e_i)}(x_j)$.

Example 5. Consider the following two GFSS where $U = \{x_1, x_2, x_3, x_4\}$ and $E = e_1, e_2, e_3$.

$$F_\rho = \begin{pmatrix} 0.2 & 0.5 & 0.9 & 1.0 & 0.6 \\ 0.1 & 0.2 & 0.6 & 0.5 & 0.8 \\ 0.2 & 0.4 & 0.7 & 0.9 & 0.4 \end{pmatrix} \quad \text{and} \quad G_\delta = \begin{pmatrix} 0.4 & 0.3 & 0.2 & 0.9 & 0.5 \\ 0.6 & 0.5 & 0.2 & 0.1 & 0.7 \\ 0.4 & 0.4 & 0.2 & 0.1 & 0.9 \end{pmatrix}$$

here

$$m(\rho, \delta) = 1 - \frac{\sum |\rho_i - \delta_i|}{\sum |\rho_i + \delta_i|} = 1 - \frac{0.1 + 0.1 + 0.5}{1.1 + 1.5 + 1.3} = 0.82$$

and

$$M_1(\widetilde{F}, \widetilde{G}) \cong 0.73, \quad M_2(\widetilde{F}, \widetilde{G}) \cong 0.43, \quad M_3(\widetilde{F}, \widetilde{G}) \cong 0.50$$

$$\therefore M_1(\widetilde{F}, \widetilde{G}) \cong 0.73$$

Hence the similarity between the two GFSS F_ρ and G_δ will be

$$S(F_\rho, G_\delta) = M_1(\widetilde{F}, \widetilde{G}) \cdot m(\rho, \delta) = 0.73 \times 0.82 \cong 0.60$$

for universal fuzzy soft sets $\rho = \delta = 1$ and $m(\rho, \delta) = 1$, then similarity $S(F_\rho, G_\delta) = 0.73$.

3 Classification Based on Soft Sets Theory

There are two concepts that underlie the classification algorithm in the soft-sets, namely classification based decision making problem as proposed by Mushrif et. al. [17], and classification algorithms based on the similarity between two fuzzy softs sets, this algorithm is a new classification algorithm proposed in this paper. We'll discuss both in this section.

3.1 Soft Sets Classifier Based on Decision Making Problems

This classifier learns by calculating the average value of each parameters (attributes or features) from all object or instant with the same class label, to construct soft sets model with universe consisting of all of class labels. In other words, an object in the universe represents all data derived from the same class label. Furthermore, to classify the test data or data of unknown class labels. First, construct the soft sets model of data using a certain formula, then construct a comparison-table in the same manner as the preparation of comparison-table in the case of decision making problem. The next step is to calculate the score to determine the class label for the data.

Mushrif et al. using this algorithm to classify the natural texture [17], we called Soft Sets Classifier (SSC). In their experiments, Mushrif et al. used 25 texture classes with 14 texture features. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less in case of soft set method in comparison with Bayes classification method.

We can use the above algorithm to classify numerical data in general. By modifying the second step in both phases of train and stage classification. To classify numerical data with this algorithm, the second step is replaced with fuzzification process, which is like counting the normalization so that all parameters have a value between 0 to 1. For example, if the classification algorithm is applied to the iris dataset. Fuzzification can be done by dividing each attributes value with the largest value at each attributes, $e_{fi} = e_i/max(e_i)$. Where e_i, $i = 1, 2, \ldots, n$ is the old attribute and e_{fi} is attribute with new value between [0,1]. We can use a different formula in this regard.

3.2 Soft Sets Classifier Based on Similarity between Two Fuzzy Soft Sets

This is a new approach to numerical classification algorithm based on the theory of fuzzy soft sets, which we propose in this paper, we refer to as Fuzzy Softs Set Classifier (FSSC). FSSC have the same learning phase with the previous soft sets classification algorithm, but the FSSC has a different classifier, it uses the similarity between two fuzzy-soft-sets. As described by Majumdar and Samanta [14]. FSSC complete algorithm is as follows.

Algorithm 6. FSSC algorithm

Pre-processing phase

1. Feature fuzzification to obtain a feature vector E_{wi}, $i = 1, 2, \ldots, N$ for all data, training dataset and testing dataset.

Training phase

1. Given N samples obtained from the data class w.
2. Calculate the cluster center vector E_w using Equation-5.

$$E_w = \frac{1}{N} \sum_{i=1}^{N} E_{wi} \tag{5}$$

3. Obtain a fuzzy soft set model for class w, $(\widetilde{F}_w, E)$, is a cluster center vector for class w having D features.
4. Repeat the process for all W classes.

Classification phase

1. Get the one unknown class data.
2. Obtain a fuzzy soft sets model for unknown class data, $(\widetilde{G}, E)$
3. Compute similarity between $(\widetilde{G}, E)$ and $(\widetilde{F}_w, E)$ for each w using Equation (4).
4. Assign the unknown data to class w if similarity is maximum.

$$w = \arg\left[\max_{w=1}^{W} S(\widetilde{G}, \widetilde{F}_w)\right]$$

If FSSC applied to the iris with fuzzification formula and data distribution (train and test) are same with the case of iris data classification in the previous examples. The results are as follows, after learning phase the fuzzy soft sets model for each class are as shown in Table-2 (a), (b), and (c).

Table 2. Tabular representasion of fuzzy soft sets model for iris dataset

(a) Setosa Class, $(\widetilde{F}_{Setosa}, E)$

$\boldsymbol{U}$	e_1	e_2	e_3	e_4
o_{setosa}	0.99	0.68	0.68	0.64

(b) Versicolor Class, $(\widetilde{F}_{Versicolor}, E)$

$\boldsymbol{U}$	e_1	e_2	e_3	e_4
o_{setosa}	0.85	0.87	0.82	0.84

(c) Virginia Class, $(\widetilde{F}_{Virginia}, E)$

$\boldsymbol{U}$	e_1	e_2	e_3	e_4
o_{setosa}	0.84	0.67	0.82	0.79

Table 3. Tabular representasion of fuzzy soft sets model for unknown class, $(\widetilde{G}, E)$

$\boldsymbol{U}$	e_1	e_2	e_3	e_4
o_{setosa}	0.63	0.52	0.48	0.40

To classify new data with the following features, $e_1 = 0.63$, $e_2 = 0.52$, $e_3 = 0.48$, and $e_4 = 0.40$, the fuzzy soft sets model $(\widetilde{G}, E)$ as shown in Table-3. First, calculate the similarity between $(\widetilde{G}, E)$ and fuzzy soft sets for each class $(\widetilde{F}_{setosa}, E)$, $(\widetilde{F}_{versicolor}, E)$, and $(\widetilde{F}_{virginia}, E)$, using Equation (4) as follows

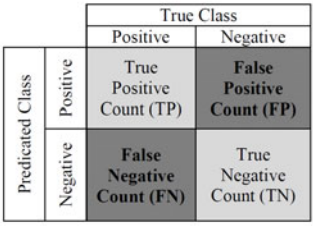
$$S_1(\widetilde{G}, \widetilde{F}_{setosa}) = 0.78,\ S_2(\widetilde{G}, \widetilde{F}_{versicolor}) = 0.89,\ S_3(\widetilde{G}, \widetilde{F}_{virginia}) = 0.79$$

In this case higest similarity is $S_2(\widetilde{G}, \widetilde{F}_{versicolor}) = 0.89$, so to the new data, we gived class label "versicolor".

The main FSSC advantage compared with previous algorithms are, have a lower complexity in the classification phase. If n is the number of test data, w is the number of class labels, and d is the number of features. So for all the test data n, the pevious algorithms required number of arithmetic operations consists of, (1) wd number of arithmetic operations to compute the soft model sets, (2) w^2d number of arithmetic operations to construct the comparison table, and (3) Finally, w^2 number of arithmetic operations to calculate the score. The total complexity for the previous algorithm is $Q(n[wd + w^2d + w^2])$. The number of arithmetic operations on the new algorithm, FSSC, for all test data n consists of (1) $2wd$ number of arithmetic operations to calculate the similarity between two fuzzy soft sets, and (2) w^2 number of arithmatic operations to seek the highest similarity value. The total complexity to FSSC is $Q(n[2wd + w^2])$. So FSSC has a lower complexity of order $Q(nw^2d)$. This order will be very influential if the number of features becomes more and more, as happened in the case of text data classification. Where the number of features can reach hundreds of thousands.

4 Classifier Evaluation

In classification problems, the primary source of performance measurements is a coincidence matrix or contingency table [18]. Figure 1. shows a coincidence matrix for a two-class classification problem. The equations of most commonly used metrics that can be calculated from the coincidence matrix is also given below



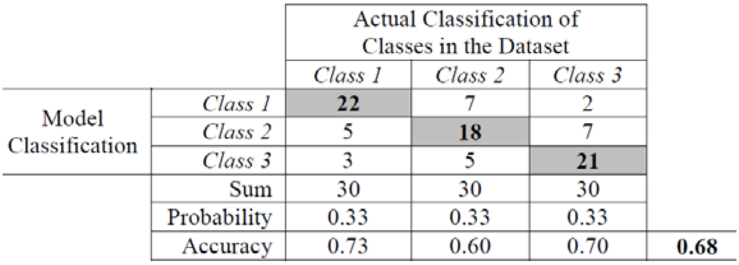
		True Class	
		Positive	Negative
Predicated Class	Positive	True Positive Count (TP)	**False Positive Count (FP)**
	Negative	**False Negative Count (FN)**	True Negative Count (TN)

Fig. 1. Contingency table for binary classification

The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples.

When the classification problem is not binary, the coincidence matrix gets a little more complicated (see Fig. 2). In this case the terminology of the performance metrics becomes limited to the "overall classifier accuracy". These formulas in Equation 6.

$$\text{(Overall Classifier Accuracy)}_i = \frac{\sum_{i=1}^{n} (True\ Classification)_i}{(Total\ Number\ of\ Cases)_i} \tag{6}$$



		Actual Classification of Classes in the Dataset			
		Class 1	*Class 2*	*Class 3*	
Model Classification	*Class 1*	**22**	7	2	
	Class 2	5	**18**	7	
	Class 3	3	5	**21**	
	Sum	30	30	30	
	Probability	0.33	0.33	0.33	
	Accuracy	0.73	0.60	0.70	**0.68**

Fig. 2. A sample coincidence matrix for a three class classifier

where i is the class number, n is the total number of classes. To minimize the effect of the imbalance between minority an mayority classes distributions, we used the weight F-measure method [2], which is equal to the harmonic mean of recall (ρ) and precision (π) [22]. Recall and precision are defined as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i}, \qquad \rho_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

Here, TP_i (True Positives) is the number of datas assigned correctly to class i: FP_i (False Positives) is the number of datas that do not belong to class i but are assigned to class i incorrectly by the classifier; and FN_i (False Negatives) is the number of datas that are not assigned to class i by the classifier but which actually belong to class i.

The F-measure values are in the interval (0,1) and larger F-measure values correspond to higher classification quality. The overall F-measure score of the entire classification problem can be computed by two different types of average, micro-average and macro-average [22].

Micro-averaged F-Measure. In micro-averaging, F-measure is computed globally over all category decisions. ρ and π are obtained by summing over all individual decisions:

$$\pi = \frac{\sum_{1=i}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FP_i)}, \qquad \rho = \frac{\sum_{1=i}^{M} TP_i}{\sum_{i=1}^{M} (TP_i + FN_i)} \tag{8}$$

where M is the number of categories. Micro-averaged F-measure is then computed as:

$$F(\text{micro-averaged}) = \frac{2\pi\rho}{\pi + \rho} \tag{9}$$

Micro-averaged F-measure gives equal weight to each data and is therefore considered as an average over all the class pairs. It tends to be dominated by the classifiers performance on common classes.

Macro-averaged F-Measure. In macro-averaging, F-measure is computed locally over each class first and then the average over all classes is taken. ρ and π are computed for each class as in Equation 7. Then F-measure for each category i is computed and the macro-averaged F-measure is obtained by taking the average of F-measure values for each category as:

$$F_i = \frac{2\pi_i\rho_i}{\pi_i + \rho_i}, \quad F(\text{macro-averaged}) = \frac{\sum_{i=1}^{M} F_i}{M} \tag{10}$$

where M is total number of classes. Macro-averaged F-measure gives equal weight to each class, regardless of its frequency. It is influenced more by the classifier's performance on rare classes. We provide both measurement scores to be more informative.

5 Experimental Results

To compare accuracy and computational time of the two soft sets classifier we perform an experiment to classify some types of classification problems, the source data comes from the UCI datasets, we select the type of classification with a real numerical features, and having multiclass, class labels more than two (binary). Accuracy is calculated using Overall Classifier Accuracy (OCA) and F-measure (micro-average and macro-average). We divide each datasets into two parts, 70% used for the training process and the remaining is for testing, and pre-process (fuzzification) applied to all datasets, to form a fuzzy value between [0,1] for each attribute. At each datasets, experiments performed at least 10 times, and train (70%) and test (30%) data were selected randomly each time we starting the experiment. The experiments are performed on a Core(TM) 2 Duo CPU, 2.1 GHz and 2 GB memory computer using MATLAB version 7.10, 32-bit.

5.1 Data Set

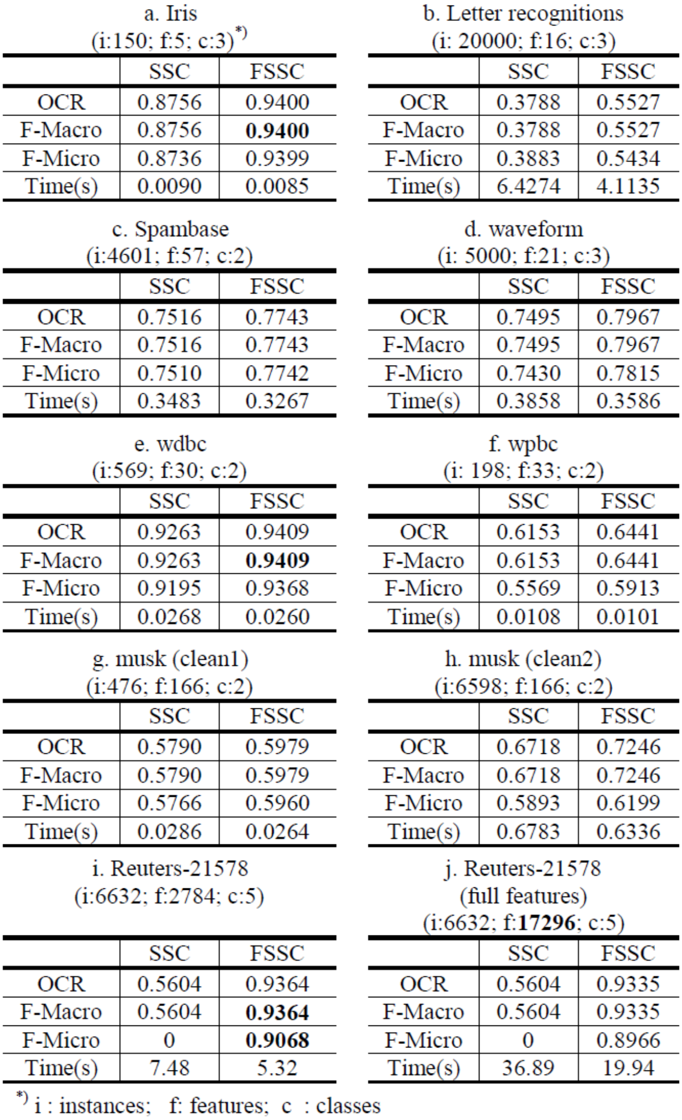
In our experiments, we used seven dataset from [6] i.e. iris, letter-recognition, breast-cancer-wisconsin (wdbc and wpdc), waveform, spambase, musk (clean1 and clean2), and Reuters-21578 document collection which had been prepared in the format matlab by [4]. Special for Reuters-21578 datasets, we used features selection algorithm used to reduce the dimension of data, in this experiment we used information gain algorithm derived from the weka [7] to select 2,784 features from 17,296 features.

5.2 Results and Discussion

From Table 4 we observe that by using seven types of dataset accuracy and time computation of the new classifier (FSSC) are always better compare to soft sets classifier proposed by Mushrif et al. (SSC). Results of experiments on classification problems with 10 types of data sets, show that in general, the classifier both can work well. The highest achievement occurred in the wdbc (Wisconsin Diagnostic Breast Cancer) dataset with accuracy (F-macros) 0.94 (FSSC) and 0.92 (SSC. While the lowest achievement occurred in the letter-recognitions dataset, ie 0.378 (SSC) and 0.55 (FSSC). In this case, the F-macros have the same value with Overall Classifier Accuracy (OCA) and we choose to represent the classifier accuracy.

Special case occurs in Reuters-21578 datasets, for this text data classification problems, with the number of attributes in the thousands and even tens of thousands, but most attribute value is 0. We used two conditions on reuters-21578 datasets, the first by using all its attributes, the number is 17,296 attributes. Second, using the attributes that have been reduced by using information gain that was reduced to 2,784 attributes. We use the five largest classes in these datasets.

Table 4. Classification accuracy and computation time for soft sets classifier proposed by Mushrif et. al. (SSC) and the new **fuzzy soft sets classifier** (FSSC)



a. Iris
(i:150; f:5; c:3)*)

	SSC	FSSC
OCR	0.8756	0.9400
F-Macro	0.8756	**0.9400**
F-Micro	0.8736	0.9399
Time(s)	0.0090	0.0085

b. Letter recognitions
(i: 20000; f:16; c:3)

	SSC	FSSC
OCR	0.3788	0.5527
F-Macro	0.3788	0.5527
F-Micro	0.3883	0.5434
Time(s)	6.4274	4.1135

c. Spambase
(i:4601; f:57; c:2)

	SSC	FSSC
OCR	0.7516	0.7743
F-Macro	0.7516	0.7743
F-Micro	0.7510	0.7742
Time(s)	0.3483	0.3267

d. waveform
(i: 5000; f:21; c:3)

	SSC	FSSC
OCR	0.7495	0.7967
F-Macro	0.7495	0.7967
F-Micro	0.7430	0.7815
Time(s)	0.3858	0.3586

e. wdbc
(i:569; f:30; c:2)

	SSC	FSSC
OCR	0.9263	0.9409
F-Macro	0.9263	**0.9409**
F-Micro	0.9195	0.9368
Time(s)	0.0268	0.0260

f. wpbc
(i: 198; f:33; c:2)

	SSC	FSSC
OCR	0.6153	0.6441
F-Macro	0.6153	0.6441
F-Micro	0.5569	0.5913
Time(s)	0.0108	0.0101

g. musk (clean1)
(i:476; f:166; c:2)

	SSC	FSSC
OCR	0.5790	0.5979
F-Macro	0.5790	0.5979
F-Micro	0.5766	0.5960
Time(s)	0.0286	0.0264

h. musk (clean2)
(i:6598; f:166; c:2)

	SSC	FSSC
OCR	0.6718	0.7246
F-Macro	0.6718	0.7246
F-Micro	0.5893	0.6199
Time(s)	0.6783	0.6336

i. Reuters-21578
(i:6632; f:2784; c:5)

	SSC	FSSC
OCR	0.5604	0.9364
F-Macro	0.5604	**0.9364**
F-Micro	0	**0.9068**
Time(s)	7.48	5.32

j. Reuters-21578
(full features)
(i:6632; f:**17296**; c:5)

	SSC	FSSC
OCR	0.5604	0.9335
F-Macro	0.5604	0.9335
F-Micro	0	0.8966
Time(s)	36.89	19.94

*) i : instances; f: features; c : classes

To this text datasets the SSC may not work well, all test data is labeled the same class. We still do not know for sure, why the SSC is not able to work well on text data, the most likely is because the text data is very sparse, very few features that have a value not equal to 0. While FSSC can still work well on these datasets, even with relatively good accuracy is 0.93. In addition, FSSC also can work faster, more and more features are used FSSC will work faster than the SSC, for full features reuters-21578 datasets, computation times for SSC is 36.89s, while FSSC only took for 19.94s. This is where one of the weaknesses of the SSC because they have to compare every feature for two different objects, when build the comparison-table, so that if the number of features becomes large, the classification process will more slowly. While FSSC does not need build the comparison-table, so it will work much faster.

6 Conclusion

In this paper we investigate the new classification based on fuzzy soft set theory. We use seven datasets from UCI to test the accuracy and computation time of the fuzzy soft sets classifier (FSSC) compares with the previous soft sets classfier proposed by Mushrif et al (SSC). In general, both can do the classifying numerical data, and both have the highest achievement in wdbc datasets, where for the FSSC accuracy is 0.94 and for SSC accuracy is 0.92. While the lowest achievement in letter-recognition classification problem, where accuracy to FSSC is 0.55 and accuracy for the SSC is 0.38. For all datasets FSSC has higher accuracy and shorter computational time. In addition to the SSC, If the number of features in the dataset the higher, then computational time become longer, also SSC can not work properly on Reuters-21578 dataset (text data). Furthermore, we will study the performance of the FSSC to classify text documents more detail, compared with a text classifier such as support vector machine (SVM), k-NN, and the others.

Acknowledgments. This work was supported by the FRGS under the Grant No. Vote 0736, Ministry of Higher Education, Malaysia.

References

1. Aktas, H., Çagman, N.: Soft sets and soft groups. Inf. Sci., 2726–2735 (2007)
2. Ali, S., Smith, K.: On learning algorithm selection for classification, App. Soft Comp. 6, 119–138 (2006)
3. Chen, D., Tsang, E., Yeung, D., Wang, X.: The parametrization reduction of soft sets and its applications. Comput. Math. Appl. 49, 757–763 (2005)
4. Cai, D., Wang, X., He, X.: Probabilistic Dyadic Data Analysis with Local and Global Consistency. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 105–112 (2009)
5. Feng, F., Jun, Y.B., Liu, X., Li, L.: An adjustable approach to fuzzy soft set based decision making J. Comp. App. Math. 234, 10–20 (2010)

6. Frank, A., Asuncion, A.: UCI Machine Learning Repository University of California, Irvine, School of Information and Computer Sciences (2010)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update (2009)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
9. Kharal, A.: Distance and Similarity Measures for Soft Sets. New Math. & Nat. Comp (NMNC) 06, 321–334 (2010)
10. Kong, Z., Gao, L., Wang, L.: Comment on a fuzzy soft set theoretic approach to decision making problems. J. Comput. Appl. Math. 223, 540–542 (2009)
11. Maji, P., Biswas, R., Roy, A.: Fuzzy soft sets. J. Fuzzy Math. 9(3), 589–602 (2001)
12. Maji, P., Roy, A., Biswas, R.: An application of soft sets in a decision making problem. Comput. Math. Appl. 44, 1077–1083 (2002)
13. Maji, P., Biswas, R., Roy, A.: Soft set theory. Comput. Math. Appl. 45, 555–562 (2003)
14. Majumdar, P., Samanta, S.: Generalised fuzzy soft sets. J. Comp. Math. App. 59, 1425–1432 (2010)
15. Majumdar, P., Samanta, S.: Similarity measure of soft sets. NMNC 04, 1–12 (2008)
16. Molodtsov, D.: Soft set theory–First results. Comp. Math. App. 37, 19–31 (1999)
17. Mushrif, M.M., Sengupta, S., Ray, A.K.: Texture classification using a novel, soft-set theory based classification algorithm. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3851, pp. 246–254. Springer, Heidelberg (2006)
18. Olson, D., Delen, D.: Advanced Data Mining Techniques, 1st edn. Springer, Heidelberg (2008)
19. Pawlak, Z.: Rough sets. Int. J. of Inform. Comput. Sci. 11, 341–356 (1982)
20. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic, Boston (1991)
21. Roy, A., Maji, P.: A fuzzy soft set theoretic approach to decision making problems. J. Comp. App. Math. 203, 412–418 (2007)
22. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM, New York (1999)
23. Zadeh, L.: Fuzzy sets. Inform. Control 8, 338–353 (1965)

An Empirical Study of Density and Distribution Functions for Ant Swarm Optimized Rough Reducts

Lustiana Pratiwi, Yun-Huoy Choo, and Azah Kamilah Muda

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Melaka, Malaysia
lustiana@gmail.com, huoy@utem.edu.my, azah@utem.edu.my

Abstract. Ant Swarm Optimization refers to the hybridization of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) algorithms to enhance optimization performance. It is used in rough reducts calculation for identifying optimally significant attributes set. Coexistence, cooperation, and individual contribution to food searching by a particle (ant) as a swarm (colony) survival behavior, depict the common characteristics of both PSO and ACO algorithms. Ant colony approach in Ant Swarm algorithm generates local solutions which satisfy the Gaussian distribution for global optimization using PSO algorithm. The density and distribution functions are two common types of Gaussian distribution representation. However, the description and comparison of both functions are very limited. Hence, this paper compares the solution vector of ACO is represented by both density and distribution function to search for a better solution and to specify a probability functions for every particle (ant), and generate components of solution vector, which satisfy Gaussian distributions. To describe relative probability of different random variables, Probability Density Function (PDF) and the Cumulative Density Function (CDF) are capable to specify its own characterization of Gaussian distributions. The comparison is based on the experimental result to increase higher fitness value and gain better reducts.

Keywords: probability density function, cumulative distribution function, particle swarm optimization, ant colony optimization, rough reducts.

1 Introduction

In the concept of Rough Sets Theory, reducts is an important attribute set which can discern all discernible objects by the original of information system (*IS*) [1] and [2]. It is the process of reducing an information system such that the set of attributes of the reduced information system is independent and no attribute can be eliminated further without losing some information from the system. In the search for relationship and global patterns in information system, it is important to be able to identify the most important attributes to represent the whole object [3], [4], [5], [6], [7], and [8]. Then the approach will set the interesting attributes that is determined by a notation of reducts.

Reducts calculation has great importance in features selection analysis. It enables the calculation of absolute reduction as well as relative reduction with core. However,

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 590–604, 2011.

the computational complexity of generating optimal reducts is very high. Since the search space increase exponentially with the number of attributes, finding the optimal reducts, a minimal reduct with minimal cardinality of attributes among all reducts is a NP-hard problem [3].

Formally, the minimum attribute reduction problem is a nonlinearly constrained combinatorial optimization problem [9]. Hence, global optimization methods can be used to solve reducts problem and gain a better result. Previous study [10] suggested to hybrid Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) as a solution for optimization problem and it serves as an enhanced process which can significantly improve the execution time, gain higher fitness values and higher predictive performance of reducts to better represent a dataset, for example in terms of classification accuracy. The PSO/ACO hybrid solution [10] uses PSO algorithm for global exploration which can effectively reach the optimal or near optimal solution. On the other hand, the ACO algorithm is used for defining a suitable fitness function to increase the competency in attribute reduction by satisfying the Gaussian distribution functions. In this paper, two types of function are compared, i.e. the Probability Density Function (PDF) and the Cumulative Density Function (CDF).

The rest of the paper is organized as follows. In Section 2, the descriptions of rough sets-based attribute reduction are presented. Section 3 explains the algorithms of ant swarm optimization technique. Section 4 discusses the fundamental theory of PDF and CDF. Section 5 briefly describes an improved rough reducts optimization framework by using PSO/ACO hybridized algorithms. The effectiveness of both functions is demonstrated, compared, and the computation results are discussed in Section 6. Finally, Section 7 outlines the conclusions followed by the future work.

2 Rough Sets-Based Attribute Reduction

Data preprocessing has become one of major point in knowledge discovery. Attribute reduction is one of the most fundamental approaches due to able to discrete some value to reduce data amounts and dimensions. In order to automate the process of reducing attributes of dataset, Rough Sets is proposed by Pawlak [3], as a mathematical technique that is applied to improve database performance by utilizing the automatically dependencies to deal better with attribute reduction without losing some information from the system [1].

In [6], Rough Set Theory [7] is still another approach to vagueness. Similarly to fuzzy set theory it is not an alternative to classical set theory but it is embedded in it. Rough set concept can be defined quite generally by means of topological operations, *interior* and *closure*, called *approximations*. Given an information system $IS = (U, A)$, where U is a non-empty finite set of objects called the universe and A is a non-empty finite set of attributes, such that $a : U \rightarrow Va$ for every $a \in A$. A reduct of A is a minimal set of attributes $B \subseteq A$ such that all attributes $a \in A - B$ are dispensable and an associated equivalence of indiscernibilty relation denoted by $IND(A) = IND(B)$. An attribute a, is said to be dispensable in $B \subseteq A$ if $IND(B) = IND(B - \{a\})$. Otherwise, the attribute is indispensable in B [1], [2], [3], [4], and [5]. Reducts are such subsets that are minimal, and do not contain any dispensable attributes. The set of all reducts of an information system IS is denoted by $RED(IS)$

or simply RED. To describe this problem more precisely, a set of objects U called the *universe* and an indiscernibility relation $R \subseteq U \times U$ are given, representing the lack of knowledge about elements of U [7]. For the sake of simplicity, R is assumed as an equivalence relation. Let X is a subset of U. The objective is to characterize the set X with respect to R. To this end, the basic concepts of rough set theory are given below

- The *lower approximation* of a set X with respect to R is the set of all objects, which can be for *certain* classified as X with respect to R.
- The *upper approximation* of a set X with respect to R is the set of all objects which can be *possibly* classified as X with respect to R.
- The *boundary region* of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not-X with respect to R.

Then the definitions of rough sets are

- Set X is *crisp* (exact with respect to R), if the boundary region of X is empty.
- Set X is *rough* (inexact with respect to R), if the boundary region of X is nonempty.

Thus a set is *rough* (imprecise) if it has nonempty boundary region; otherwise the set is *crisp* (precise) [6]. The approximations and the boundary region can be defined more precisely. To this end some additional notation are needed. The equivalence class of R determined by element x will be denoted by $R(x)$. The indiscernibility relation in certain sense describes the lack of knowledge about the universe. Equivalence classes of the indiscernibility relation, called *granules* generated by R, represent elementary portion of knowledge able to perceive due to R. Thus in view of the indiscernibility relation, in general, it is possible to observe individual objects but is forced to reason only about the accessible granules of knowledge [7]. Formal definitions of approximations and the boundary region are as follows

- *R-lower approximation* of X

$$R_*(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}. \tag{1}$$

- *R-upper approximation* of X

$$R^*(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \varnothing\}. \tag{2}$$

- *R-boundary region* of X

$$RN_R(X) = R^*(X) - R_*(X). \tag{3}$$

The definition approximations are expressed in terms of granules of knowledge. The lower approximation of a set is union of all granules which are entirely included in the set; the upper approximation – is union of all granules which have non-empty intersection with the set; the boundary region of set is the difference between the upper and the lower approximation. Rough sets can be also defined employing, instead of approximation, rough membership function [8].

$$\mu_X^R : U \to < 0,1 >, \text{ where } \mu_X^R(x) = \frac{| X \cap R(x) |}{| R(x) |} \tag{4}$$

and |*X*| denotes the cardinality of *X*. The rough membership function expresses conditional probability that *x* belongs to *X* given *R* and can be interpreted as a degree that *x* belongs to *X* in view of information about *x* expressed by *R* [8]. The rough membership function can be used to define approximations and the boundary region of a set, as shown below [8]

$$R_*(X) = \{x \in U : \mu_X^R(x) = 1\}, \tag{5}$$

$$R^*(X) = \{x \in U : \mu_X^R(x) > 0\}, \tag{6}$$

$$RN_R(X) = \{x \in U : 0 < \mu_X^R(x) < 1\}. \tag{7}$$

There are two definitions of rough sets, which are as follow

Definition 1: Set *X* is *rough* with respect to *R* if $R_*(X) \neq R^*(X)$.

Definition 2: Set *X rough* with respect to *R* if for some *x*, $0 < \mu_X^R(x) < 1$.

The definition 1 and definition 2 are not equivalent [6] and rough set theory clearly distinguishes two very important concepts, vagueness and uncertainty, very often confused in the Artificial Intelligence literature. Vagueness is the property of sets and can be described by approximations, whereas uncertainty is the property of elements of a set and can be expressed by the rough membership function.

3 Ant Swarm Optimization Algorithms

Ant Swarm Optimization refers to the hybridizarion of both PSO and ACO algorithms to solve optimization problem. Both PSO and ACO algorithms adapt swarm intelligence metaheuristics which is based on population global search and co-operative biologically inspirited algorithm motivated by social analogy [10]. PSO was inspired by real life social behavior of bird flocking or fish schooling, while ACO imitates foraging behavior of real life ants. PSO still has the problems of dependency on initial point and parameters, difficulty in finding their optimal design parameters, and the stochastic characteristic of the final outputs for local searching [10].

On the other hand, ACO has positive feedbacks for rapid discovery of good solutions and a simple implementation of pheromone-guided will improve the performance of other optimization techniques, for example in [11], the simulation results has shown that combination of ACO with Response Surface Method (RSM), can be very successively formulating an optimized minimum surface roughness prediction model for reduction of the effort and time required. And thus in this study, a simple pheromone-guided mechanism is explored to increase the performance of PSO method for rough reducts optimization [10].

3.1 Particle Swarm Optimization (PSO)

In PSO, particles as candidate solutions of a population, simultaneously coexist and evolve based on knowledge sharing with neighboring particles. Each particle generates a solution using directed velocity vector, while flying through the problem search space. Each particle modifies its velocity to find a better solution (position) by applying its own flying experience for the best position memory found in the earlier flights and experience of neighboring particles as the best-found solution of the population [10]. Each particle's movement is the composition of an initial random velocity and two randomly weighted influences; individuality, the tendency to return to the particle's best position P_{best}, and sociality, the tendency to move forwards the best previous position of the neighborhoods G_{best}.

Particles update their positions and velocities as shown below

$$v_{t+1}^{i} = w_t v_t^i + c_1 r_1 (p_t^i - x_t^i) + c_2 r_2 (p_t^g - x_t^i), \tag{8}$$

$$x_{t+1}^{i} = x_t^i + v_{t+1}^{i}, \tag{9}$$

where x_t^i represents the current position of particle i in solution space and subscript t indicates an iteration count; p_t^i is the best-found position of particle i up to iteration count t and represents the cognitive contribution to the search velocity v_t^i . Each component of v_t^i can be clamped to the range $[-v_{max}, v_{max}]$ to control excessive roaming of particles outside the search space; p_t^g is the global best-found position among all particles in the swarm up to iteration count t and forms the social contribution to the velocity vector; r_1 and r_2 are random numbers uniformly distributed in the interval (0, 1), while c_1 and c_2 are the cognitive and social scaling parameters, respectively; w_t is the particle inertia, which is reduced dynamically to decrease the search area in a gradual fashion by Shi et al. in [12]. The variable w_t is updated along with the iterations in (10)

$$w_t = (w_{max} - w_{min}) * \frac{t_{max} - t}{t_{max}} + w_{min}, \tag{10}$$

where, w_{max} and w_{min} denote the maximum and minimum of w_t respectively, verifying from 1.4 to 0.4; t_{max} is a given number of maximum iterations. Particle i flies toward a new position according to (8) and (9). In this way, all particles P of the swarm find their new positions and apply these new positions to update their individual best p_t^i points and global best p_t^g of the swarm. This process is repeated until iteration count $t = t_{max}$ (a user-defined stopping criterion is reached).

Given an information system $IS = (U, A)$, $A = (C \cup D)$, where C is a non-empty finite set of condition attributes and D is a non-empty finite set of decision attributes, such that $RED \subseteq C$. Attributes of PSO consists of P denotes the number of particles in the population; $f(x_t^i)$ represents the objective function value of particle i at position x and calculated as

$$f(x_t^i) = \alpha * \gamma_{x_t^i}(D) + \beta * \frac{|C| - |x_t^i|}{|C|} , \tag{11}$$

where $\gamma_{x_t^i}(D)$ is the classification quality of particle condition attribute set x_t^i, which contains the reducts RED, and relative to decision table D, defined as follows

$$\gamma_{x_t^i}(D) = \frac{r_{RED}}{r_C}, \tag{12}$$

where r_{RED} represents a degree of dependency of RED on D and r_C represents a degree of dependency of C on D. $|x_t^i|$ is the '1' number of the length of selected feature subset or the number of attributes for particle x_t^i, while population of solutions P is at iteration count t. $|C|$ is the total number of condition attributes. α and β parameters correspond to importance of classification quality and subset length, $\alpha \in [0,1]$ and $\beta = 1 - \alpha$.

PSO techniques can also generate high-quality solutions within shorter calculation time and stable convergence characteristics than other stochastic methods [13]. PSO has shown fast convergence speed and global search ability [15]. However, Suganthan in [12] mentioned the major drawback of PSO, like in other heuristic optimization techniques, is that it has a slow fine tuning ability of solution quality. PSO is also a variant of stochastic optimization techniques, which is requiring relatively a longer computation time than mathematical approaches. PSO still has the problems of dependency on initial point and parameters, difficulty in finding their optimal design parameters, and the stochastic characteristic of the final outputs [15].

Shi and Eberhart proved that PSO converges fast under all cases but will slow its convergence speed down when reaching the optima [12]. This may be due to the use of the linearly decreasing inertia weight. By using the linearly decreasing inertia weight, the PSO is lacking of global search ability at the end of run, even when the global search ability is required to jump out of the local minimum in some cases [16].

3.2 Ant Colony Optimization (ACO)

ACO studies the concept of "the emergent collective intelligence of groups of simple agents". As [17] discussed ACO algorithm was motivated by ants' social behavior. Ants have no sight and are capable of finding the shortest route between a food source and their nest when moving from one place to another. Ants deposit substance, called pheromone on the ground to form a trail by using strong pheromone concentrations to decide shorter paths [17].

ACO is particularly attractive for feature selection since there is no heuristic information that can guide the search to the optimal minimal subset every time. Besides, if features are represented as a graph, ants can discover the best feature combinations as they traverse the graph [17]. Dre´o, et al. and Socha in [10] found that ants behave as social insects that directly more toward the survival of the colony as a whole than that of a single individual of the colony. Indirect co-operative foraging process of ants is very interesting behavior to be adopted in searching problem of PSO.

The ants also use their capability to locate their food resources found by their mates and proven those behaviors to stimulate the optimization of ant foraging behavior in ACO. The implementation of ACO in the Ant Swarm approach was based on the studies by Angeline in [10], which has been proven that PSO was able to discover reasonable quality solutions much faster than other evolutionary algorithms. However, PSO does not possess the ability to improve upon the quality of the solutions as the number of generations is increased and this problem can be improved by ACO.

4 Probability Density and Cumulative Distribution Functions

Each continues random variable X has an associated Probability Density Function (PDF) $f(x)$ [18]. It "records" the probability associated with X as areas under its graph. More precisely, "the probability that a value X is between a and b" $= P(a \leq X \leq b) = \int_a^b f(x)dx$.

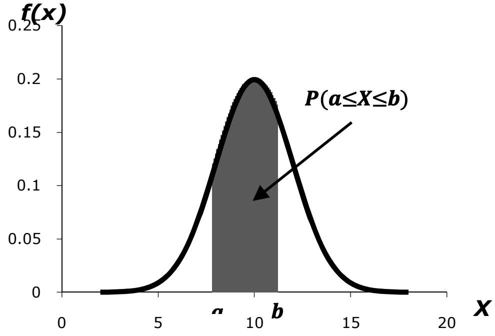


Fig. 1. Areas under a probability density function f on the interval $[a,b]$[18]

That is, the probability that X takes on a value in the interval $[a,b]$ is the area under the density function from a to b.

For example [18],

$$P(1 \leq X \leq 3) = \int_1^3 f(x)dx, \tag{13}$$

$$P(3 \leq X) = P(3 \leq X < \infty) = \int_3^\infty f(x)dx, \tag{14}$$

$$P(X \leq -1) = P(-\infty < X \leq -1) = \int_{-\infty}^{-1} f(x)dx. \tag{15}$$

i) Since probabilities are always between 0 and 1, it must be that $f(x) \geq 0$, so that $\int_a^b f(x)dx$ can never give a "negative probability", and

ii) Since a "certain" event has probability 1, $P(-\infty < X < \infty) = 1 = \int_{-\infty}^{\infty} f(x)dx =$ total area under the graph of $f(x)$.

The properties i) and ii) are necessary for a function $f(x)$ to be the PDF for some random variable X. In 1809, Gaussian distribution was applied to the first application of normal distribution as a central role in probability theory and statistics [18]. Normal distribution is an important tool to approximate the probability distribution of the average of independent random variables. A continues random variable has a normal distribution with parameters μ and $\sigma^2 > 0$ if its probability density function f is given by [19]

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \text{ for } -\infty < x < \infty. \tag{16}$$

This distribution function is denoted as [19]

$$f(x) = \mathcal{N}(p_t^g, \sigma). \tag{17}$$

There is another function, The Cumulative Distribution Function (CDF) which records the cumulative distribution function same probabilities associated with, but in a different way. CDF records the same probabilities associated with X, but in a different way [18]. The CDF $F(x)$ is defined by

$$F(x) = P(X \leq x). \tag{18}$$

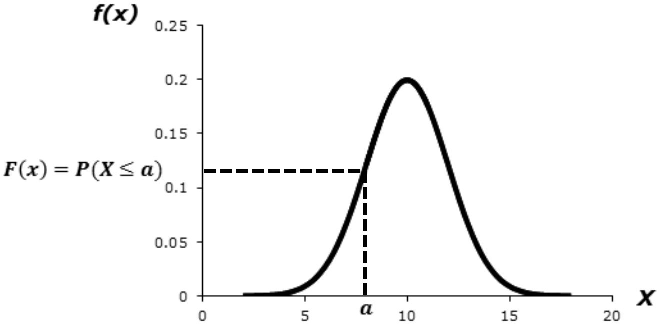


Fig. 2. A graphical representation of the relationship between PDF and CDF [18]

$F(x)$ gives the "accumulated" probability "up to x" and be able to be seen immediately how PDF and CDF are related and defined by [19]

$$F(x) = P(X \leq x) = \int_{\infty}^{x} f(t)dt. \tag{19}$$

Notice that $F(x) \geq 0$, since it's probability, and that [17]

a) $\lim_{x\to\infty} F(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{\infty} f(t)dt = 1$ and

b) $\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{-\infty} f(t)dt = 0$, and that

c) $F'(x) = f(x)$ (by the Fundamental Theorem of Calculus).

Items c) states the connection between the CDF and PDF in another way, which is the CDF $F(x)$ is a derivative of the PDF $f(x)$ and therefore if X has a $\mathcal{N}(p_t^g, \sigma)$ distribution, then its distribution function is given by [19]

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{1}{2}[1 + \operatorname{erf}(\frac{x-\mu}{\sigma\sqrt{2}})] \text{ for } -\infty < x < \infty. \tag{20}$$

5 Ant Swarm Optimization for Rough Reducts (ASORR)

This paper proposed to solve the rough reducts optimization problem with an improved particle swarm optimization hybridized with an ant colony approach, called PSO/ACO [10]. The method applied PSO for global optimization and the idea of ACO approach to update positions of particles to attain rapidly the feasible solution space.

The term ''reduct'' corresponds to a wide class of concepts. What typifies all of them is that they are used to reduce information (decision) systems by removing irrelevant attributes. Given an information system $IS = (U, A)$, a reduct is a minimal set of attributes $B \subseteq A$ such that $IND(B) = IND(A)$, where $IND(B)$, $IND(A)$ are the indiscernibility relations defined by B and A, respectively by Z. Pawlak and A. Skowron in [20]. The intersection of all reducts is called a core. Intuitively, a reduct is a minimal set of attributes from A that preserves the original classification defined by A. Reducts are extremely valuable in applications. Unfortunately, finding a minimal reduct is NP-hard in the general case. One can also show that, for any m, there is an information system with m attributes having an exponential number of reducts. Fortunately, there are reasonably good heuristics which allow one to compute sufficiently many reducts in an acceptable amount of time [20].

The implementation of the algorithm consists of two stages [13]. In the first stage, it applies PSO, while ACO is implemented in the second stage (see Fig. 2). ACO works as a local search, wherein, ants apply pheromone-guided mechanism to update the positions found by the particles in the earlier stage. The implementation of ACO in the second stage of Ant Swarm is based on the studies by Angeline in [10] which showed that PSO discovers reasonable quality solutions much faster than other evolutionary algorithms. Thus, PSO does not possess the ability to improve upon the quality of the solutions as the number of generations is increased. Ants also use their capability to locate their food resources found by their mates and proved those behaviors to stimulate the optimization of ant foraging behavior in ACO. In the Ant Swarm approach, a simple pheromone-guided search mechanism of ant colony is implemented which acts locally to synchronize positions of the particles of PSO to quickly attain the feasible domain of objective function [10].

Consider a large feature space full of feature subsets [10]. Each feature subset can be seen as a point or position in such a space. If there are N total features, then there will be 2^N kinds of subset, different from each other in the length and features contained in each subset. The optimal position is the subset with least length and highest classification quality. Now a particle swarm is put into this feature space, each particle takes one position. The particles fly in this space, their goal is to fly to the best position [15]. Over time, they change their position, communicate with each other, and search around the local best and global best position. Eventually, they should converge on good, possibly optimal, positions. It is this exploration ability of particle swarms that should better equip it to perform feature selection and discover optimal subsets [10].

5.1 Representation of Position and Velocity

The particle's position is represented as binary bit strings of length N, where N is the total number of attributes. Every bit represents an attribute, the value '1' means the corresponding attribute is selected while ‘0’ not selected. Each position is an attribute subset [10]. The velocity of each particle is represented as a positive integer, varying between 1 and V_{max}. It implies how many of the particle’s bits (features) should be changed, at a particular moment in time, to be the same as that of the global best position, i.e. the velocity of the particle flying toward the best position [10]. The number of different bits between two particles relates to the difference between their positions [10].

For example, $P_{gbest} = [1\,0\,1\,1\,1\,0\,1\,0\,0\,1]$, $p_t^i = [0\,1\,0\,0\,1\,1\,0\,1\,0\,1]$. The difference between $Gbest_t$ and the particle’s current position is $P_{gbest} - p_t^i = [1 - 1\,1\,1\,0 - 1\,1\ -1\,0\,0]$. A value of 1 indicates that compared with the best position, this bit (feature) should be selected but is not, which will decrease classification quality and lead to a lower fitness value. Assume that the number of 1’s is a. On the other hand, a value of -1 indicates that, compared with the best position, this bit should not be selected, but is selected. Irrelevant features will make the length of the subset longer and lead to a lower fitness value. The number of -1’s is b. The value of $(a - b)$ is used to express the distance between two positions; $(a - b)$ may be positive or negative.Such variation makes particles exhibit an exploration ability within the solution space. In this example, $(a - b) = 4 - 3 = 1$, so $p^g - p_t^i = 1$ [10].

5.2 Representation of PDF and CDF in ACO

In the application of Ant Swarm approach, a simple pheromone-guided search mechanism of ant colony was implemented which acted locally to synchronize positions of the particles in PSO to attain the feasible domain of the objective function [10] and [21] faster. The proposed ACO algorithm from the previous research [10] handles P ants as equal to the number of particles in PSO. Each i ant generates a solution, z_t^i around the global best-found position among all particles in the swarm, p_t^g up to the iteration count, t as

$$z_t^i = \mathcal{N}(p_t^g, \sigma). \tag{21}$$

In (14), the algorithm generates components of solution vector z_t^i, which satisfy Gaussian distributions as according to (10). This distribution function has properties to determine the probabilities of z_t^i, which are mean $\mu = p_t^g$ and standard deviation σ, where, initially at $t = 1$ value of $\sigma = 1$ and is updated at the end of each iteration as $\sigma = \sigma \times d$, where, d is a parameter in (0.25, 0.997) and if $\sigma < \sigma_{min}$ then $\sigma = \sigma_{min}$, where, σ_{min} is a parameter in (10^{-2}, 10^{-4}). In this stage, PDF and CDF were applied and compared using the same components as above. Density function will adapt the probabilities associated areas with random variables in (9). Otherwise, distribution function will calculate the accumulative probabilities up to the same random variables of PDF in (13). And then according to each of function computation of z_t^i in (8), evaluate objective function value $f(z_t^i)$ using z_t^i in (4) and replace position x_t^i the current position of particle i in the swarm if $f(z_t^i) < f(x_t^i)$ as $x_t^i = z_t^i$ and $f(x_t^i) = f(z_t^i)$ [10]. This simple pheromone-guided mechanism considers, there is highest density or distribution of trails (single pheromone spot) at the global best solution p_t^g of the swarm at any iteration $t + 1$ in each stage of ACO implementation and all ants

P search for better solutions in the neighborhood of the global best solution [10] and this process is repeated until iteration $t = t_{max}$.

In the beginning of the search process, ants explore larger search area in the neighborhood of p_t^g due to the high value of standard deviation $\sigma$and intensify the search around p_t^g. Thus, ACO not only helps PSO to efficiently perform global exploration for rapidly attaining the feasible solution space but also to effectively reach the optimal fitness value to gain better reducts, as the algorithm progresses as the following algorithms [13]:

```
Step 1: Initialize Optimization
Initialize algorithm constants $t_{max}$, $P$, and $\{0,1\}^m$ is the
  $m$-dimensional Boolean particle space.
Calculate the inertia weight of each particle space in
  (10).
Initialize randomly all particle positions $x_t^i$ and
  velocities $v_t^i$.
Initialize the positive acceleration constants $c_1, c_2$ and
  $MaxFit$ as the maximum fitness value.

Step 2: Perform Optimization (Initialization)
Do {
 For each particle {
  Evaluate objective function value $f(x_t^i)$ in (11)
  Assign best position to each particle $p_t^i$ = $x_t^i$ with
    $Pbest_t^i = f(x_t^i)$, $i$ = 1,$\cdots$,$P$
  Evaluate $f_t^{best}(p_t^{best}) = min\{Pbest_t^1, Pbest_t^2, \cdots, Pbest_t^P\}$
  If is $Pbest_t^i$ better than the best fitness value
    $f_t^{best}(p_t^{best})$ in history {
   Assign current fitness value as $Gbest_t = f_t^{best}(p_t^{best})$
   Assign $p_t^g$ = $p_t^{best}$
  } End
} While maximum iterations

Step 3: Perform Optimization (Update the positions and
        velocities)
Do {
 For each particle {
  Update particle position $x_t^i$ and velocity $v_t^i$ according
    (8) and (9) to all $P$ particles
  Evaluate objective function value $f(x_t^i)$ in (11)
  Generate $P$ solutions $z_t^i$ using (13)
  Evaluate objective function value $f(z_t^i)$ in (11)
  If $f(z_t^i)$ is less than $f(x_t^i)$ {
   Assign  $f(z_t^i)$ to $f(x_t^i)$ and $z_t^i$ to $x_t^i$
  }
  If $Pbest_t^i$ greater than $f(x_t^i)$ {
        Update $Pbest_t^i = f(x_t^i)$ and $p_t^i = x_t^i$
      }
      Evaluate $f_t^{best}(p_t^{best}) = min\{Pbest_t^1, Pbest_t^2, \cdots, Pbest_t^P\}$
      If is $Pbest_t^i$ better than the best fitness value
         $f_t^{best}(p_t^{best})$ in history {
        Assign current fitness value as $Gbest_t = f_t^{best}(p_t^{best})$
        Assign $p_t^g = p_t^{best}$
      } End
    Return subset of attributes
  } While (maximum iterations and $Gbest_t < MaxFit$)
```

Step 4: Report best solution p^g as P_{gbest} global best position of the swarm with objective function value $f(p^g)$

6 Experimental Result

The performance of the proposed enhanced Ant Swarm algorithm for global optimization function has been tested on several well-unknown benchmark multimodal problems [13]. All the test functions are multimodal in nature. Because of the characteristics, it is difficult to seek for the global minima. Particle Swarm Ant Colony Optimization (PSACO) algorithm parameter settings used in all the simulations is given as: number of particles, P = 10; cognitive and social scaling parameters, $c1$ = 2, $c2$ = 2; maximum and minimum values of inertia weights, w_{max} = 0.7, w_{min} = 0.4; maximum number of iterations, t_{max} = 100 * n, n is the size of solution vector. Implementation of ASORR has been tested on14 datasets. The experimental results are reported based on the number of reducts and iterations, fitness values, and the classification accuracy for performance analysis are shown in Table 1.

Both PDF and CDF algorithms implementations use Naïve Bayes to extract rules from the data for rule induction in classification. Ten-fold cross validation was applied to estimate the classification accuracy.

The two types of enhanced Ant Swarm algorithms are compared and the best solutions of both algorithms found are presented. Experimental results have shown that PDF has more optimal results than CDF in most of datasets. Table 1 reports the result produced by PDF in a number of reducts evaluation is smaller than CDF but in average, both function yield the same results. However, a number of reducts is not able to determine which function devises better result, in terms of finding the optimal fitness value and gain higher classification accuracy. The fitness values of each function reported in Table 1 has shown that PDF can achieve better values and prove that a smaller number of reducts will increase the fitness value. As shown in Table 1, PDF achieved better results than CDF by reducing 5.19 iterations for number of iteration results analysis in average for 10 independent runs. Table 1 shows the results in terms of classification accuracy where PDF takes more significant optimal solution and has the same analysis results with its fitness value performances, and gain better accuracy than CDF.

Table 1. ASORR Experimental Results on No. of Reducts, Fitness Value, No. of Iteration, and Classification Accuracy (%)

No.	Dataset	Features	Instances	Criteria										
				No. of Reducts		Fitness Value			No. of Iteration			Classification Accuracy		
				ASORSAR		ASORSAR		Best Result	ASORSAR		Best Result	ASORSAR		Best Result
				CDF	PDF	CDF	PDF		CDF	PDF		CDF	PDF	
1	Soybean-small	35	47	3	3	0.9014	0.9057	PDF	86	78	PDF	97.98	97.55	CDF
2	Lung	56	32	5	5	0.9036	0.9161	PDF	90	77	PDF	73.28	74.69	PDF
3	Zoo	16	101	4	4	0.6848	0.6890	PDF	100	100	None	94.8	95.3	PDF
4	Lymphography	18	148	5	5	0.6646	0.6645	CDF	100	100	None	78.31	78.51	PDF
5	Corral	6	64	4	4	0.3333	0.3333	None	100	100	None	84.38	84.38	None
6	Vote	16	300	5	5	0.6255	0.6237	CDF	100	100	None	92.42	92.73	PDF
7	DNA	57	318	6	6	0.8884	0.8949	PDF	100	100	None	28.74	29.04	PDF
8	M-of-N	13	1000	6	6	0.5385	0.5385	None	100	100	None	95.7	95.7	None
9	Exactly	13	1000	6	6	0.5385	0.5385	None	100	100	None	68.8	68.8	None
10	Exactly2	13	1000	10	10	0.2308	0.2308	None	100	100	None	75.8	75.8	None
11	Led	24	2000	6	5	0.7688	0.7813	PDF	100	100	None	100	100	None
12	Mushroom	22	8124	3	3	0.8301	0.8343	PDF	100	100	None	78.31	79.56	PDF
13	Breastcancer	9	699	3	2	0.6507	0.6619	PDF	100	100	None	95.24	95.09	CDF
14	Tic-Tac-Toe	9	958	6	6	0.2558	0.2574	PDF	100	100	None	69.24	69.19	CDF
Average				5	5	0.6296	0.6336	PDF	98	97	PDF	80.93	81.17	PDF

Thus, based on the results obtained and presented, the implementation of ASORR using PDF is having better performances, both in gaining higher fitness value and better quality of reducts as compared to CDF. However, some previous studies [7, 18] also have explored Ant Swarm algorithm in other implementation field and all of them applied PDF as solution for the optimal results.

7 Conclusion

An extensive comparative study on rough reducts optimization has been presented. This paper compared the merits of PDF and CDF focusing on PSO/ACO enhanced rough reducts. A simple-pheromone-guided mechanism is implemented as local search by using Gaussian distribution functions, the PDF and the CDF, to improve the performance of PSO algorithm. The experiments have shown that PDF is better than CDF in terms of generating smaller number of reducts, improved fitness value, lower number of iterations, and higher classification accuracy. Thus, the experimental results have also provided further justification on previous studies which have implemented PDF instead of CDF into Ant Swarm algorithm in various domains. The initial results of PDF are promising in most of the tested datasets. Hence, future works are to test on the enhanced ASORR algorithm with PDF in various domains to validate its performance in yielding the optimal reducts set.

Acknowledgement

This work is funded by the Ministry of Higher Education Malaysia and Universiti Teknikal Malaysia Melaka (UTeM) through the Fundamental Research Grant Scheme, FRGS/2009/FTMK(1)-F0072.

References

1. Pawlak, Z.: Some issues on rough sets. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B.z., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 1–58. Springer, Heidelberg (2004)
2. Wang, X., Xu, R., Wang, W.: Rough Set Theory: Application in Electronic Commerce Data Mining. In: Proc.s of the IEEE/WIC/ACM Intl. Conf. on Web Intelligence. IEEE Computer Society, Los Alamitos (2004)
3. Barceló, Juan, A.: Computational Intelligence in Archaeology. 1599044897. Idea Group Inc., IGI (2008)
4. Wang, J., Meng, G., Zheng, X.: The Attribute Reduce Based on Rough Sets and SAT Algorithm. Shanghai: IEEE. Intelligent Information Technology Application 1 (2008)
5. Pawlak, Z., Skowron, A.: Rudiments of Rough Sets. Information Science 177, 3–27 (2007)
6. Pawlak, Z.: Rough Sets. University of Information Technology and Management, Poland (2004)
7. Pawlak, Z.: Rough Sets. Int. J. of Information and Computer Sciences. 11, 341–356 (1982)

8. Pawlak, Z., Skowron, A.: Rough Membership Function. In: Yeager, R.E., Fedrizzi, M., Kacprzyk, J. (eds.) Advances in the Dempster-Schafer of Evidence, pp. 251–271. New York (1994)
9. Ye, D., Chen, Z., Ma, S.: A new fitness function for solving minimum attribute reduction problem. In: Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS, vol. 6401, pp. 118–125. Springer, Heidelberg (2010)
10. Shelokar, P.S., et al.: Particle Swarm and Ant Colony Algorithms Hybridized for Improved Continues Optimization. In: Applied Mathematics and Computation, vol. 188(1), pp. 129–142. Elsevier Inc., Amsterdam (2007)
11. Kadirgama, K., Noor, M.M., Alla, A.N.A.: Response Ant Colony Optimization of End Milling Surface Roughness. J.s of Sensors. 10, 2054–2063 (2010)
12. Shi, Y., Eberhart, R.C.: Empirical study of particle swarm optimization. In: IEEE. Int. Congr. Evolutionary Computation, vol. 1, pp. 101–106 (2001)
13. Wang, X., et al.: Feature Selection Based on Rough Sets and Particel Swarm Optimization, vol. 28(4). Elsevier Science Inc., New York (2007)
14. Lee, K.Y., Park, J.B.: Application of Particle Swarm Optimization to Economic Dispatch Problem: Advantages and Disadvantages. In: PSCE, vol. 1, IEEE, Los Alamitos (2006)
15. Yang, Qin and Wang, Danyang.: An Improved Particle Swarm Optimization. In: IEEE, BMEI (2009)
16. Salehizadeh, S.M.A., et al.: Local Optima Avoidable Particle Swarm Optimization. In: Swarm Intelligence Symposium, pp. 16–21. IEEE, Nashville (2009)
17. Chen, Y., Miao, D., Wang, R.: A Rough Set Approach to Feature Selection Based on Ant Colony Optimization, vol. 31(3). Elsevier, Amsterdam (2010) Springer (2007)
18. Freiwald, R.: The Cumulative Distribution Function for a Random Variable X. Washington University, `http://www.math.wustl.edu/~freiwald/Math132/cdf.pdf` (retrieved on March 15, 2011)
19. The Probability Density and Cumulative Distribution Functions. ReliaSoft, `http://www.weibull.com/LifeDataWeb/the_probability_density_and_cumulative_distribution_functions.htm` (retrieved on March 14, 2011)
20. Pawlak, Z., Skowron, A.: Rough Sets and Boolean Reasoning, vol. 177, pp. 41–73. Elsevier, Amsterdam (2007)
21. Kaveh, A., Talatahari, S.: A Hybrid Particle Swarm and Ant Colony Optimization for Design of Truss Structures. Asian J. of Civil Engineering (Building and Housing) 9(4), 329–348 (2008)

The UTLEA: Uniformization of Non-uniform Iteration Spaces in Three-Level Perfect Nested Loops Using an Evolutionary Algorithm

Shabnam Mahjoub[1] and Shahriar Lotfi[2]

[1] Islamic Azad University-Shabestar Branch
Shabnam.mahjoub@yahoo.com
[2] Computer Science Department, University of Tabriz
Shahriar_lotfi@tabrizu.ac.ir

Abstract. The goal of the uniformization based on the concept of vector decomposition, to find the basic dependence vector set in a way that any vector in iteration space could present non-negative integer combination of these vectors. To get an optimal solution, we can use an approximate algorithm. In this paper, the uniformization for three-level perfect nested loops has been presented using an evolutionary method that is called the UTLEA, the method to minimize both the number of vectors and dependence cone size. The most available approaches have not been used; moreover, there are problems in approaches that could generalize them in three levels. In the proposed approach, we have been tried to solve these problems and according to executed tests, the achieved results are close to optimal result.

Keywords: Uniform and Non-uniform Iteration Space, Vector Decomposition, Uniformization, Loop Parallelization, Evolutionary Algorithm.

1 Introduction

A challenging problem for parallelizing compilers is to defect maximum parallelism [8]. According to the studies [12], most of the execution time of computational programs is spent in loops. Since then parallelizing compilers have focused on loop parallelism. In fact, parallelizing compiler to get the parallel architectural advantages generates parallel code in a way that generated code had the dependence constraints in that program. Then the iterations of the loop can be spread across processors by having different processors executing different iterations simultaneously. One of the simplest approaches used for parallelism is WaveFront which in all of the iterations in the same WaveFront are independent of each other and depend only on the iteration in the previous WaveFront [13], [18]. From this point, we can find out the importance of uniformizations. Dependence constraints in iteration loop are known as cross-iteration dependence. The loop with no cross-iteration dependence is known as *doall* which could execute in any order. Simply parallelizing of these loops is possible. But if a loop had cross-iteration dependence, known as *doacross*, parallelizing of these loops is very harder than previous loops [19], [20].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 605–617, 2011.

There are several methods to deal with nested loops. We might break the dependences and change the loop into the other loop that didn't have any cross-iteration dependences. If it is not possible, we can still execute the loop in parallel in a way that proper synchronization had added to impose cross-iteration dependences. If all techniques fail, the *doacross* loop must be executed serially [4].

There are three major difficulties in parallelizing nested loops [21]. First, to enter correct simultaneity, compilers or programmers have to find out all cross-iteration dependences. But until now there has not been any dependences analysis method that could efficiently identify all cross-iteration dependences unless in a condition that dependence pattern would be uniform for all the iterations. Second, although all the cross-iteration dependences can been identified, it is difficult to systematically arrange synchronization primitives especially when the dependence pattern is irregular. Finally, the synchronization overhead which will significantly degrade the performance should be minimized.

In this paper, a new method using an evolutionary approach has been presented for uniformization of non-uniform iteration space of a three-level nested loop that is called UTLEA and has been analyzed after executing on different dependences. The rest of the paper is organized as follows; in section 2 the problem is explained, in section 3 the basic concepts for a better understanding are explained, in section 4 related work, in section 5 the proposed method and in section 6 evaluation and experimental results are explained.

2 The Problem

In general, loops with cross-iteration dependences are divided in two groups. First group is loops with static regular dependence which can be analyzed during compile time and the second group is loops with dynamic irregular dependences. The loops of the second group for the lack of sufficient information can not be parallelized in the compile time. To execute such loop efficiently in parallel, runtime support must be provided. Major job of parallelizing compilers is to parallelize the first group loops. These loops are divided into two subgroups. Loops with uniform dependences and loops with non-uniform dependences. The dependences are uniform when the patterns of the dependence vectors are uniform. In other words, the dependence vectors have been expressed by constants or distance vectors. But if dependence vectors in irregular patterns have not been expressed by distance vector, these are known as non-uniform dependences [10].

Parallelizing nested loops have several stages. These are involving data dependence analysis, loop tilling [16], loop generation and loop scheduling [1], [6], [7], [17], [18]. The uniformization is performed in the data dependence analysis stage. The result of this step is dependence vectors between loop iterations that are possible to have non-uniform pattern. To facilitate generating parallel code with basic vectors, this non-uniform space changes to uniform space. The goal of doing this job is to decrease basic dependence vector sets (BDVSs) in a new space. Although the dependence cone size (DCS) of the basic vectors should be minimum, to seek small and simple set of uniform dependence vectors, to cover all of the non-uniform dependences in the nested loop. Then the set of basic dependences will be added to every iteration to replace all original dependence vectors.

3 Background

In this section, necessary basic concepts have been presented.

3.1 Dependence Analysis

Common methods to compute data dependence is to solve a set of equations and inequalities with a set of constraints which are the iteration boundaries. In the result, two methods presented for solving the dependence convex hull (DCH). Both of which are valid. In simple cases, using one set of these solutions is sufficient. But in complex cases, in which dependence vectors are very irregular, we have to use both sets of solutions. These two DCH represented in one DCH as the complete DCH (CDCH) and is proved that DCDH includes complete information about dependences [10].

3.2 Dependence Cone and Dependence Cone Size

For the dependence vector set D, the dependence cone C(D) is defined as the set [4], [5]:

$$C(D) = \{\bar{x} \in R^n : \bar{x} = \lambda_1 \bar{d}_1 + ... + \lambda_m \bar{d}_m, \lambda_1, ..., \lambda_m \geq 0\}. \tag{1}$$

And the DCS, assuming each d_i that means the DCS is defined as the area of the intersection of $d_1^2+d_2^2+\ldots d_n^2=1$ with dependence cone C(D). In fact, dependence cone is the smaller cone that includes all of the dependence vectors of the loop. In three-level space, DCS is proportional to enclosed volume between the basic dependence vectors and sphere with r = 1.

3.3 Evolutionary Algorithm Overview

Darwin's gradual evolution theory has been inspiring source for evolutionary algorithms. These algorithms are divided into five branches which genetic algorithm is special kind of those. Using genetic algorithm [9] for optimum process was proposed by Holland in 1975. Inventing this algorithm as an optimization algorithm has been on base of simulating natural development and it has been based on the hefty mathematical theory. Developing optimization process is on the base of random changes of various samples in one population and selecting the best ones. Genetic algorithm as an optimal computational algorithm efficiently seeks different areas of solution space considering a set of solution space pointes in any computational iteration. Since all of the solution spaces have been sought, in this method, against one directional method, there will be little possibility for convergence to a local optimal point. Other privilege of this algorithm needs determining objective value in different points and do not use other information such as derivative function. Therefore, this algorithm could be used in various problems such as linear, nonlinear, continuous and discrete.

In this algorithm, each chromosome is indicative on a point in solution space. In any iteration, all of available chromosome are decoded and acquired objective function. Based on the stated factors each chromosome has been attributed fitness. Fitness will determine selection probability to each chromosome, and with this selection probability, collections of chromosome have been selected and the new chromosomes will be generated applying genetic operator on them. These new chromosomes will be replaced with previous generated chromosomes. In executing this algorithm, we need to 4 parameters such as generation size, initial population size, crossover rate and mutation rate.

4 Related Works

The first method called naive decomposition [4]. Although a simple and clear method, a contradiction might exist in parallel execution of iterations.

Tzen and Ni [21] proposed the dependence uniformization technique based on solving a system of Diophantine equations and a system of inequalities. In this method, maximum and minimum of dependence slops have been computed according to dependence pattern for two-level iteration spaces. Then by applying the idea of vector decomposition, a set of basic dependences is chosen to replace all original dependence constraints in every iteration so that the dependence pattern becomes uniform. But since one of the vectors (0, 1) or (0, -1) should be in BDVS, DCS remains large.

In first method of Chen and Chung yew [3], the maximum number of basic vectors is three. In this method there are several selections without limitation for BDVS and any selection has different efficiency. Thus, this method needs a selected strategy that chooses a set which decreases the synchronization overhead and increases the parallelism. Therefore, in the second method of them [4] has been tried to improve the proposed method which in DCS is close to original DCS from non-uniform dependences and in fact this method has been improved to minimize the DCS in two-levels.

Chen and Shang [5] have proposed three methods on the basis of three possible measurements in a way that goal of any method is achieving to maximum of that measure. This method can be used for three-level spaces but the direction of dependence has not been considered. Furthermore, the optimization for the DCS which greatly affects parallelism has not been studied.

In the method according to evolutionary approach, genetic algorithm for two-level spaces has been used to solve the problem [15]. Although acquired results are not certain, but are very close to optimal solution for two-level spaces.

In general, in recent years in the area of uniformization a minority studies have been done. Therefore, it is required to perform better optimization.

5 The UTLEA Method

In this paper, the genetic algorithm is used as an instrumentation and searching method for finding basic vectors in uniformization of non-uniform three-level iteration spaces. In the following, different stages of proposed method are presented.

5.1 Coding

In the proposed method, every chromosome indicated a BDVS. Since the nested loop is three-level, chromosomes are spotted as a three-dimensional array. Every gene of chromosomes involves three x, y and z components that all of them indicate one vector in three-level space. In the following, an example of chromosome is shown. U_1, U_2 and U_3 are upper bound for three-level nested loop. Because of the loop index variables are integer, the amounts of x_i, y_i and z_i are integer. Also, the reason of selecting y_i between $-s_1$ and s_1 and z_i between $-s_2$ and s_2 is for equaling the amounts on the bases of loop execution.

x_1	x_2	x_3
y_1	y_2	y_3
z_1	z_2	z_3

$$0 \le x_i \le U_1$$

$$-s_1 \le y_i \le s_1, \quad s_1 = \lfloor U_1 / 2 \rfloor$$

$$-s_2 \le z_i \le s_2, \quad s_2 = \lfloor U_2 / 2 \rfloor$$

Fig. 1. Coding in the UTLEA method

5.2 Objective and Fitness Function

In this problem, a minimization problem, because of maximum nature of fitness function, it is on the contrary of objective function. In general, three factors are playing roles in determining fitness of every chromosome and we should consider these factors one by one.

Length of chromosomes. In this paper, the length of every chromosome is shown by L(i) function in which i is the chromosome number. Because the goal is to minimize length of BDVS, the 1/L(i) statement is added to fitness function. Since, the length of any chromosomes is not zero, L(i) is never zero and thus 1/L(i) will not take undefined value. At the beginning of implementation of algorithm, the length of all chromosomes is equal to 5. But because the optimal length is between n and 2n-1, during the implementing of algorithm, chromosomes have variable lengths are between 3 and 5.

Computing DCS(i) of chromosomes. In this paper, DCS is shown by DCS(i) in which i is the chromosome number. 1/DCS(i) statement like the least length is added to fitness function. Since, DCS(i) could have zero, the 1/DCS(i) statement could take undefined value. For solving this problem, we could have 1/(DCS(i)+1) statement instead of 1/DCS(i). Fig. 1 showing the dependence cone in three-level space.

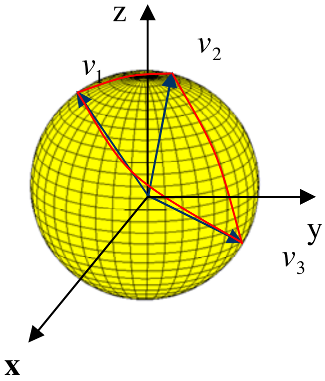


Fig. 2. DCS in three-level spaces

It is clear, if the coordinate system is changed, the considering DCS is not changed. Therefore, for computing this volume, we could change the coordinate system in a way that one of the vectors coincides with a main axis. For doing this, the vectors should rotate in a way that one of them coincides with z axis. A used rotation matrix is a trigonometric and clockwise matrix shown in relation 2. We can compute the considering volume after the rotation using a triple integral. It is better instead of Cartesian coordinates, spheral coordinates can be used to determine upper bounds and lower bounds of this integral. After rotating, the φ for the vector that coincides with z axis is equal to zero. Also we can consider the ρ for each of these vectors equal to one. Therefore in this section, calculation of φ and θ for two other vectors after rotating, is sufficient. Finally by using relation 3, we could compute DCS for chromosomes with L(i)=3.

$$R = \begin{bmatrix} \cos(\alpha).\cos(\beta) & -\sin(\alpha) & \cos(\alpha).\sin(\beta) \\ \sin(\alpha).\cos(\beta) & \cos(\alpha) & \sin(\alpha).\sin(\beta) \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \tag{2}$$

$$\begin{cases} DCS(i) = \int_{\theta_1}^{\theta_2} \int_0^{f(\theta)} \int_0^1 \rho^2 \sin\varphi \,\mathrm{d}\rho \,\mathrm{d}\varphi \,\mathrm{d}\theta \\ f(\theta) = \dfrac{\Delta\varphi}{\Delta\theta}(\theta - \theta_1) + \varphi_1 \end{cases} \tag{3}$$

In relation 2, variable α indicates the angle of the projection vector in 2-D Cartesian space x and y with positive direction of x axis and variable β indicates the angle of vector with positive direction of z axis. In relation 3, $f(\theta)$ is a linear interpolation function. Although, more complex functions [14] can be used to calculate the volume, the same linear function is sufficient.

The above method just could compute the DCS for basic vectors with L(i)=3. For longer lengths, we can use other method or universalize this method in a way that is used for longer than three. Even if all vectors are located in convex space of three vectors in three-level space, we can use this method. But there is a problem which in general in three-level space, we can not say that all vectors are located in convex space of three vectors. As a whole, if we consider the weight of every gene of chromosome equal to one number and the found result of three-vector method to be in the form of DCS3(a, b, c) function, in which a, b, c are the weight of vectors, in this case we can use the following method for computing the BDVS for chromosomes with length 4 or 5.

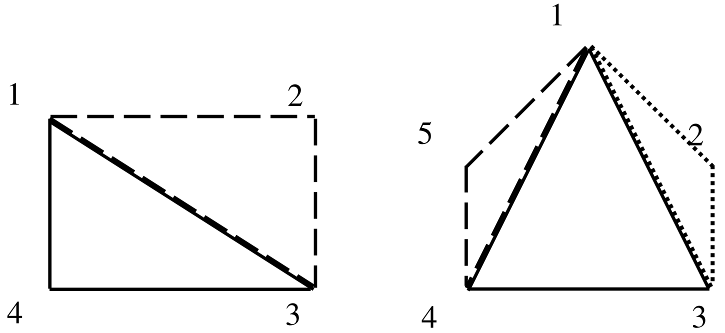


Fig. 3. Computing DCS for chromosomes with L(i)>3

$$DCS(i) = DCS3(1,\ 2,\ 3) + DCS3(1,\ 3,\ 4),\ if\ L(i) = 4 \tag{4}$$

$$DCS(i) = DCS3(1,\ 2,\ 3) + DCS3(1,\ 3,\ 4) + DCS3(1,\ 4,\ 5),\ if\ L(i) = 5 \tag{5}$$

For using the above method for lengths longer than three, all vectors in chromosomes should be sorted correctly. For this reason, all of the vectors have been rotated until a vector could coincide with z axis. This vector's number is 1. Then for other vectors, their projection angle in 2-D Cartesian space x and y with negative direction of x axis are computed and are used as a weight to sort the vectors. The weight of each vector is considered between 0 and 360 in order to prevent different vectors to have equal length. In fact, the vector with negative y element, the number of 360 is detracted of occurred absolute value.

Computing M(i) of chromosomes. In this paper, the number of construable dependence vectors extent by BDVS of chromosome i are shown as M(i) function. By solving Diophantine equation [2], [10], [11] in relation 6 the amount of M(i) are acquired for chromosome i.

$$\alpha_1(x_1,\ y_1,\ z_1) + \ldots + \alpha_n(x_n,\ y_n,\ z_n) = (x,\ y,\ z) \tag{6}$$

In relation 6 α_j is as unknown of equation that must be non-negative integer, then basic vectors will not have any problem in executing of the loop. In fact, relation 6 showed that whether dependence vector (x, y, z) are decomposable by basic vectors of chromosome i or not. Therefore, fitness function is computed as the following:

$$f(i) = (w_1 / L(i)) + (w_2 /(DCS(i)+1)) + (w_3 \times M(i)) \tag{7}$$

w_j are weights considered for showing the importance of function. Here w_1=1, w_2=3 and w_3=2.

5.3 Selection Operator

The selection operator applied in this minimization problem is tournament selection operator.

5.4 Crossover Operator

The crossover operator applied in this problem is 2-point crossover with different cutting points. These points are shown in fig. 4 by cpoint_1 and cpoint_2 for chromosome 2i-1 and cpoint_3 and cpoint_4 for chromosome 2i. If the length of chromosomes after crossover operation is longer than 5, these genes of the chromosome are eliminated.

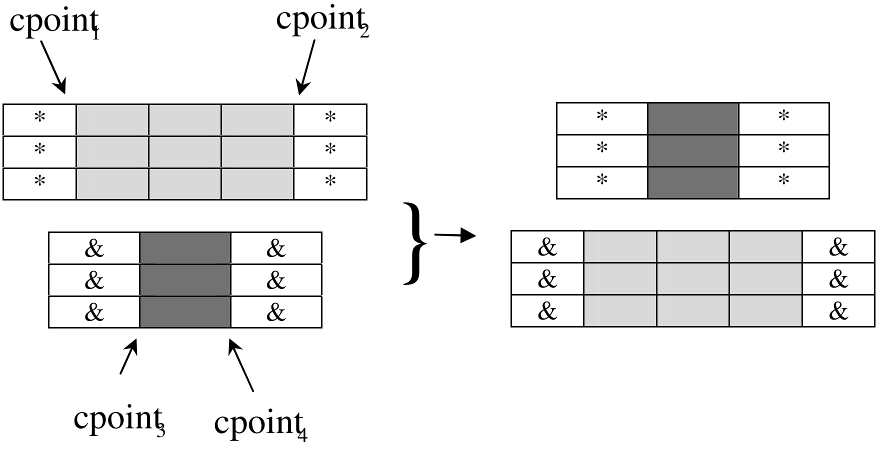


Fig. 4. Two-point crossover operation with different cutting points

5.5 Mutation Operator

In the mutation operator applied in this problem, one gene of chromosome is selected randomly and then its amount is replaced with one possible amount of other. In fig. 5 an example of mutation operation is shown.

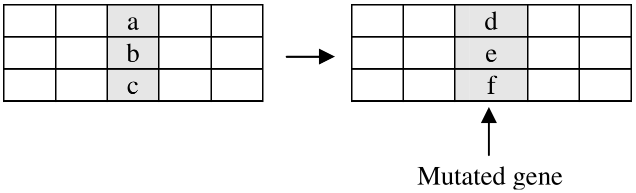


Fig. 5. Mutation operation

By doing these operators on the chromosomes may generate infeasible chromosomes. For this reason, penalty technique has been used.

$$f(i)=f(i)-0.5\,f(i),\ \textit{if}\ \ M(i)=0 \tag{8}$$

$$f(i)=f(i)-0.4\,f(i),\ \textit{if there is an invalid gene such as}\ (-1,\ 0,\ 0\,),\ \textit{etc.} \tag{9}$$

Also, there is an elitism operator before selection operator that means in every generation, the best chromosome is selected and transformed to intermediate generation directly.

6 Evaluation and Experimental Results

The UTLEA is executed in vb6 and according to various dependences, some results of its performance are presented in this section.

6.1 Experimental Results

To verify the proposed method, many tests have been executed. In this section, the results of three tests are summarized in Table 1.

Table 1. The results of the UTLEA

Dependence vectors	DCS	Fitness	Result
Uniform iteration space with vectors (1, 2, 1), (1, 2, 3) and (3, 2, 1)	0.114	21.025	{(1, 2, 1), (3, 2, 1), (0, 0, 1)}
Uniform iteration space with vectors (1, 2, -2), (1, -2, 2) and (0, 1, 1)	0.741	26.055	{(1, 2, -2), (1, -2, 2), (0, 1, 1)}
Uniform iteration space with vectors or combination of them (0, 1, -1), (0, 0, 1) and (1, 0, 0)	0.718	82.088	{(0, 1, -1), (0, 0, 1), (1, 0, 0)}

The stability of these tests is shown in fig. 6.

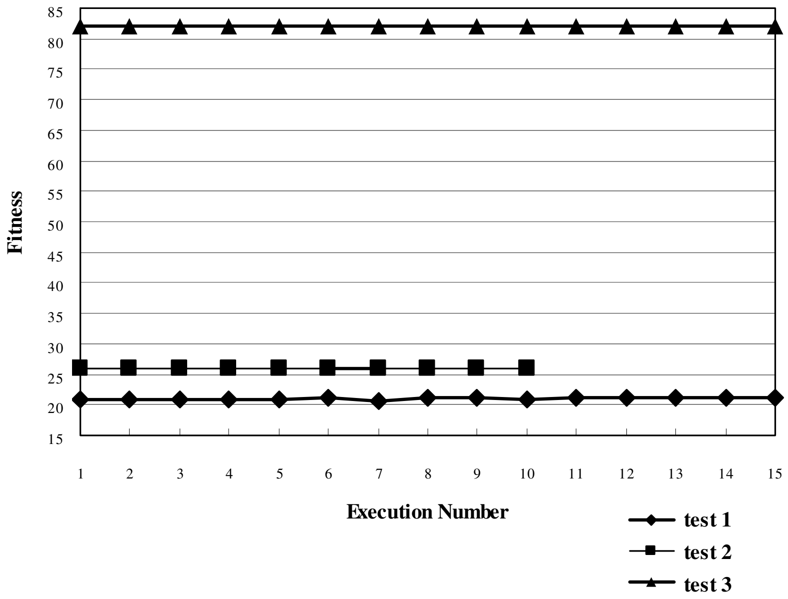


Fig. 6. Stability of the UTLEA

For showing the convergence of the UTLEA, these tests are executed again with the parameters summarized in table 2 which k is selection parameter, pc is crossover rate, pm is mutation rate, init_pop is initial population size and num_gene is the number of generation.

Table 2. The prameters of the tests

Tests	K	Pc	Pm	Init_pop	Num_gene	Result
Test 1	4	0.7	0.01	1000	1000	{(1, 2, 1), (3, 2, 1), (0, 0, 1)}
Test 2	4	0.7	0.01	500	500	{(1, 2, -2), (1, -2, 2), (0, 1, 1)}
Test 3	4	0.7	0.01	500	500	{(0, 1, -1), (0, 0, 1), (1, 0, 0)}

The convergence of these tests is shown in fig. 7, 8 and 9 respectively.

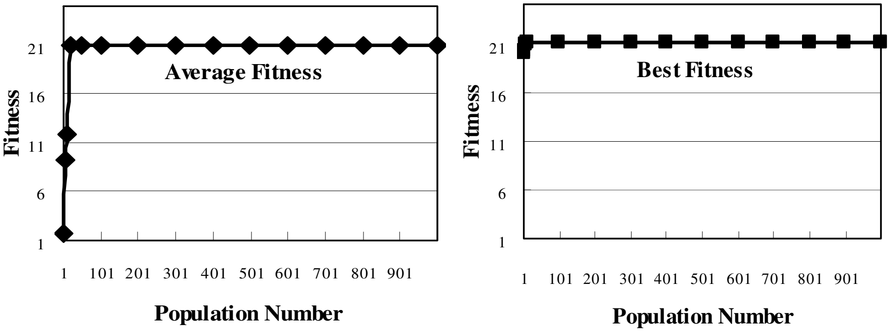


Fig. 7. Convergence of test 1

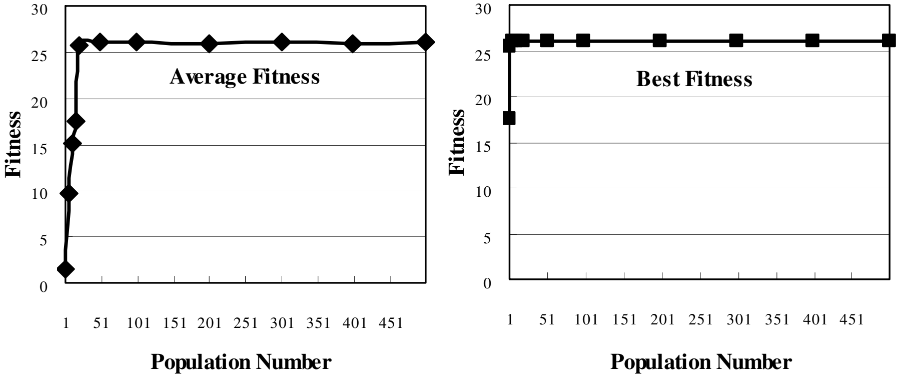


Fig. 8. Convergence of test 2

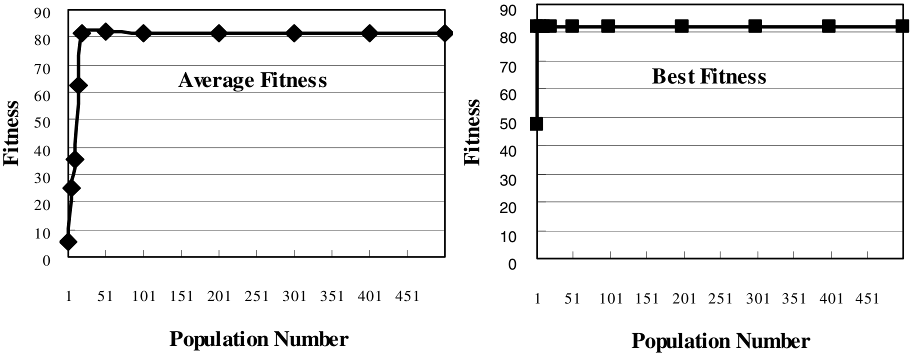


Fig. 9. Convergence of test 3

In addition, other tests have been executed on a larger scale. For example, this algorithm for 500 dependences of uniform spaces with vectors {(1, 2, 1), (2, -1, 1), (2, 1, -1)} is executed. The best result for this test is {(1, 2, 1), (2, -1, 1), (2, 1, -1)} with DCS=0.267 and fitness=1000.699.

6.2 Comparison with Other Methods

Comparison with other methods is summarized in table 3. For example, the last test (test 4) that its code is given in this section is executed to compare the proposed method with other methods. The result of following test according to basic area 1 in Chen and Shang's method [4] is {(1, 0, 0), (0, 1, 0), (0, 0, 1)} with DCS=0.523, according to basic area 2 is {(1, 2, 0), (-2, -4, 1)} with DCS=0 but x component is negative and according to basic area 3 is {(1, 2, 1), (1, 0, 0), (0, 1, 0), (0, 0, 1)} with DCS=0.523. But the result from executing the UTLEA has obtained {(2, 3, -3), (1, 2, 1), (1, 2, 0)} with DCS=0.0087. Also, the result of the UTLEA for this test is better than Chen and Yew's method [3].

```
For i=1 To 15
    For j=1 To 15
        For k=1 To 15
            A=(3i+j-1, 4i+3j-3, k+1)=…
            …=A(i+1, j+1, k)
        End For
    End For
End For
```

Table 3. Comparison the UTLEA method with other methods

Uniformization methods	The number of basic vectors	Considering the direction vectors	Other descriptions
Chen and Shang	Unknown (the most cases is large)	×	The DCS remains large
Tzen and Ni	Unusable in three-level spaces	×	-
Chen and Yew	Most of the time is large in three-level (5)	✓	here are at least a main vector in three-level spaces
The Method based on evolutionary approach	Unusable in three-level spaces	✓	-
The UTLEA	Between 3 and 5	✓	The DCS is small and the number of basic vectors is optimal

7 Conclusion and Future Works

In this paper, a dependence uniformization method is presented by using an evolutionary approach for three-level non-uniform iteration spaces called the UTLEA. Most of the previous methods only used in two-level spaces. In some of the methods used in three-level have not been paid attentions to direction of vectors. In other words, outcome basic vectors were not acceptable considering execution of loops. Also, the uniformization algorithm according to evolutionary approach presented for two-level spaces, are not applicable in three-level spaces too, since the dependence cone size in three-level spaces is different from two-level. Therefore, in this paper we have tired to solve the defects of previous methods and do a correct uniformization for three-level spaces.

As future works, a method based on evolutionary approach for uniformization of two and three levels together is suggested and the genetic parameters used in this proposed method can be improved.

References

1. Andronikos, T., Kalathas, M., Ciorba, F.M., Theodoropoulos, P., Papakonstantinou, G.: An Efficient Scheduling of Uniform Dependence Loops. Department of Electrical and Computer Engineering National Technical University of Athens (2003)
2. Banerjee, U.: Dependence Analysis for Supercomputing. 101 Philip Drive, Assinippi Park, Norwell, 02061. Kluwer Academic, Massachusetts (1988)
3. Chen, D., Yew, P.: A Scheme for Effective Execution of Irregular DOACROSS Loops. In: Int'l Conference on Parallel Processing (1993)
4. Chen, D.K., Yew, P.C.: On Effective Execution of Non-uniform DOACROSS Loops. IEEE Trans. On Parallel and Distributed Systems (1995)
5. Chen, Z., Shang, W., Hodzic, E.: On Uniformization of Affine Dependence Algorithms. In: 4 the IEEE Conference on Parallel and Distributed Processing (1996)
6. Darte, A., Robert, Y.: Affine-By-Statement Scheduling of Uniform Loop Nests Over Parametric Domains. J. Parallel and Distributed Computing (1995)
7. Darte, A., Robert, Y.: Constructive Methods for Scheduling Uniform Loop Nests. IEEE Trans. Parallel Distribut. System (1994)
8. Engelmann, R., Hoeflinger, J.: Parallelizing and Vectorizing Compilers. Proceedings of the IEEE (2002)
9. Goldberg, D.E.: Genetic Algorithm in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)
10. Ju, J., Chaudhary, V.: Unique Sets Oriented Parallelization of Loops with Non-uniform Dependence. In: Proceedings of International Conference on Parallel Processing (1997)
11. Kryryi, S.L.: Algorithms for Solving of Linear Diophantine Equations in Integer Domains. Cybernetics and Systems Analysis, 3–17 (2006)
12. Kuck, D., Sameh, A., Cytron, R., Polychronopoulos, A., Lee, G., McDaniel, T., Leasure, B., Beckman, C., Davies, J., Kruskal, C.: The Effects of Program Restructuring, Algorithm Change and Architecture Choice on Program Performance. In: Proccedings of the 1984 International Conference on Parallel Processing (1984)
13. Lamport, L.: The Parallel Execution of DO Loops. Comm. ACM 17(2), 83–93 (1974)
14. Murphy, J., Ridout, D., Mcshane, B.: Numerical Analysis Algorithms and Computation. Ellis Harwood, New York (1995)
15. Nobahari, S.: Uniformization of Non-uniform Iteration Spaces in Loop Parallelization Using an Evolutionary Approach. M. Sc. Thesis, Department of Computer Engineering (2009) (in Persian)
16. Parsa, S.: A New Genetic Algorithm for Loop Tilling. The Journal of Supercomputing, 249–269 (2006)
17. Parsa, S.: Lotfi, Sh.: Parallel Loop Generation and Scheduling. The Journal of Supercomputing (2009)
18. Parsa, S., Lotfi, S.: Wave-front Parallelization and Scheduling. In: 4th IEEE International Conference on Parallel Processing, pp. 382–386 (2007)
19. Sawaya, R.: A Study of Loop Nest Structures and Locality in Scientific Programs. Maste of Applied Science Graduate Department of Electrical and Computer Engineering University of Torento (1998)
20. Tzen, T.H.: Advance Loop Parallelization: Dependence Uniformization and Trapezoid Self Scheduling. Ph. D. thesis, Michigan State University (1992)
21. Tzen, T.H., Ni, L.: Dependence Uniformization: A loop Parallelization Technique. IEEE Trans. on Parallel and Distributed Systems. 4(5), 547–558 (1993)

A Discrete Event Simulation for Utility Accrual Scheduling in Uniprocessor Environment

Idawaty Ahmad, Shamala Subramaniam,
Mohamad Othman, and Zuriati Zulkarnain

Faculty of Computer Science and Information Technology,
University Putra Malaysia UPM
Serdang 43400 Selangor Malaysia
idawaty@fsktm.upm.edu.my

Abstract. This research has focused on the proposed and the development of an event based discrete event simulator for the existing General Utility Scheduling (GUS) to facilitate the reuse of the algorithm under a common simulation environment. GUS is one of the existing TUF/UA scheduling algorithms that consider the Time/Utility Function (TUF) of the executed tasks in its scheduling decision in a uniprocessor environment. The scheduling optimality criteria are based on maximizing accrued utility accumulated from execution of all tasks in the system. These criteria are named as Utility Accrual (UA). The TUF/ UA scheduling algorithms are design for adaptive real time system environment. The developed GUS simulator has derived the set of parameter, events, performance metrics and other unique TUF/UA scheduling element according to a detailed analysis of the base model.

Keywords: Time/Utility Function, Real Time Scheduling, Discrete Event Simulation and Uniprocessor.

1 Introduction

Real-time scheduling is basically concerned with satisfying a specific application time constraints. In adaptive real time system an acceptable deadline misses and delays are tolerable and do not have great consequences to the overall performances of the system.

One of the scheduling paradigms in adaptive real time system environment is known as Time/Utility Function (TUF)[1],[2]. A TUF of a task specifies the quantified value of utility gained by the system after the completion of a task shown in Fig. 1. The urgency of a task is captured as a deadline on X-axis and the importance of a task is measured by utility in Y-axis. With reference to Fig. 1, in the event of the task being computed at time A, which denotes the range between the start of execution and the stipulated deadline, the system gains a positive utility.

However, if the task is completed at time B, which causes failure of deadline compliance requirement, the system acquires zero utility.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 618–626, 2011.

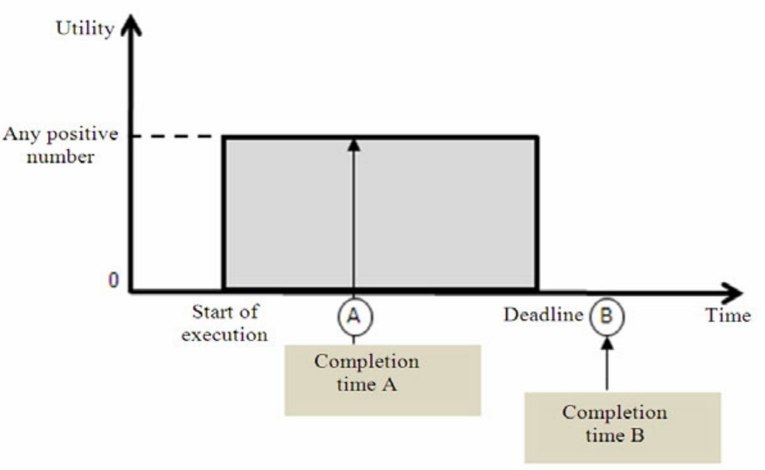


Fig. 1. Time/Utility Function

1.1 Problem Statement

GUS is a uniprocessor TUF/UA scheduling algorithm that manages the independence tasks and tasks that have dependencies with other tasks [3], [4]. The dependencies are due to the sharing of resources via the single unit of resource request model. In enhancing and developing the GUS algorithm, performance analysis and its respective tools are evident.

Though there exists the simulation tools, there does not exist a detailed description and a developed General Purpose Language (GPL) DES for the TUF/UA scheduling domain specifically the GUS algorithm. The lack of uniformity in the choice of simulation platforms is a clear limitation for investigating the performances of the TUF/UA scheduling algorithms.

1.2 Objective

The objective of this research is to build the GUS simulator from the scratch to enable customization requirements of any research and to provide the freedom to understand, configure TUF modules, draw desired scheduling environment and plot the necessary performance graphs. In order to evaluate and validate the performance of the designed simulator, a simulation model for the TUF/UA scheduling environment is deployed.

2 A Discrete Event Simulation Framework

A discrete event simulation framework is developed to verify the performance of the GUS scheduling algorithm. In order to precisely remodel and further enhance the GUS algorithm, DES written in C language in Visual C++ environment is the best method to achieve this objective. Fig.2 shows the developed GUS simulator framework. It consists of the four major components i.e., the DES simulation, scheduling algorithm, entities and resources components.

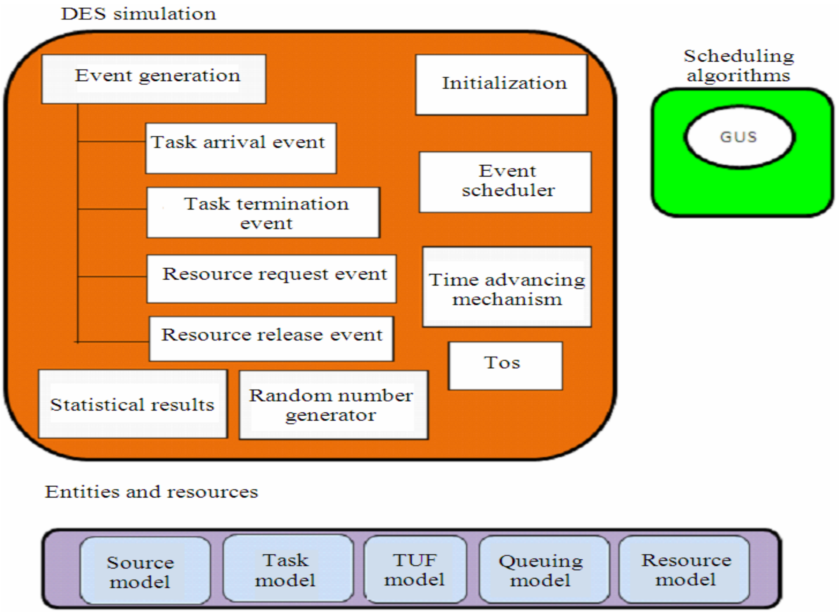


Fig. 2. A Disrete Event Simulation Framework in Uniprocessor Environment

The core component to execute the developed simulator consists of the events, events scheduler, time advancing mechanism, random number generator, Termination Of Simulation (TOS) and statistical results as depicted in Fig. 2. A flow chart of the execution of the simulator is depicted in Fig. 3. It illustrates the structure of the simulation program and the events involved. The initialization triggers the deployment of the entire simulation. Relating the norm of an idle system, no task can depart without invoking its creation (i.e., Task Arrival Event). Thus, the assumption of the event arrival schedule is set to 0.0000.

Referring to Fig. 3, after initialization the next pre-requisite mandatory step is to scan the event list and select the event with the earliest time of occurrence. Mapping the selection to DES is embedded in the time advancing mechanism (i.e., simulation clock). The simulation clock is then advanced to the time of occurrence of the selected event. The simulator then executes the selected event and updates the system state variables affected by the event.

Each of the identified events is auctioned by calling an associated event routine which results in the addition of future events to the event list. The execution of event routines is done to achieve the stipulated two purposes to model the deployment of an event and to track the resource consumption status of the event. Referring to Fig. 3, the defined events and their respective routine descriptions in this research are the task arrival, resource request, resource release and task termination event. The completion of the simulation will be done upon the convergence of the repetitive structure to a predefined value which also known as TOS. TOS is critical in determining the validity of the acquired results. It must represent the system in entirety. In this research, the simulator is terminated of two conditions i.e., the event list is empty and the arrival of task termination event for the final task in the simulator is executed.

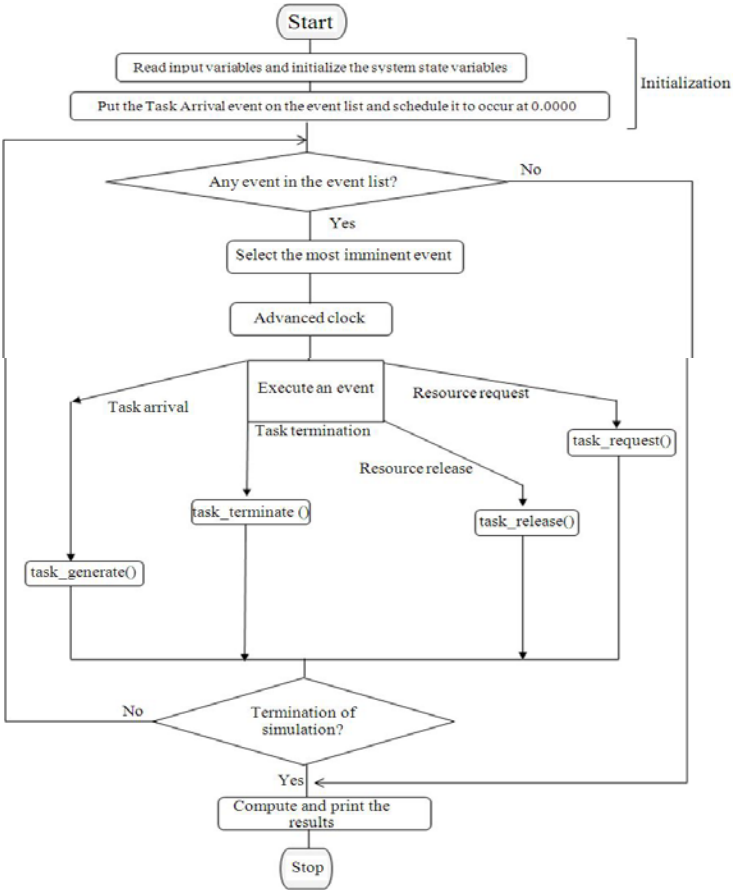

Start
Read input variables and initialize the system state variables
Put the Task Arrival event on the event list and schedule it to occur at 0.0000
Initialization
Any event in the event list?
No
Yes
Select the most imminent event
Advanced clock
Execute an event
Task arrival
Task termination
Resource request
Resource release
task_request()
task_terminate ()
task_release()
task_generate()
No
Termination of simulation?
Yes
Compute and print the results
Stop

Fig. 3. Flowchart of the Simulation Program

Entities are the representation of objects in the system [6]. Fig. 4 shows the interaction between the entities and resource models that are designed throughout the simulator. i.e., the source and tasks entities, the resources and a queue of an unordered task list named as *utlist*.

Simulating the source model involves the representation of the load generation of the system under study. It is vital to accurately represent the load to ensure the algorithms deployed are tested on the actual scenario. A source injects a stream of tasks into the system. The maximum numbers of tasks are 1000 and denoted as MAX_TASKS.

A micro abstraction of a source is the task model. Each task is associated with an integer number, denoted as tid. Each task is associated with an integer number, denoted as tid. Upon generation, a task is executed for 0.50 seconds (i.e., the average execution time denoted as C_AVG). Given the task average execution time C_AVG and a load factor load, the tasks inter arrival time follows exponential distribution with mean value of C_AVG/load. Fig.5 shows a task as a single flow of execution.

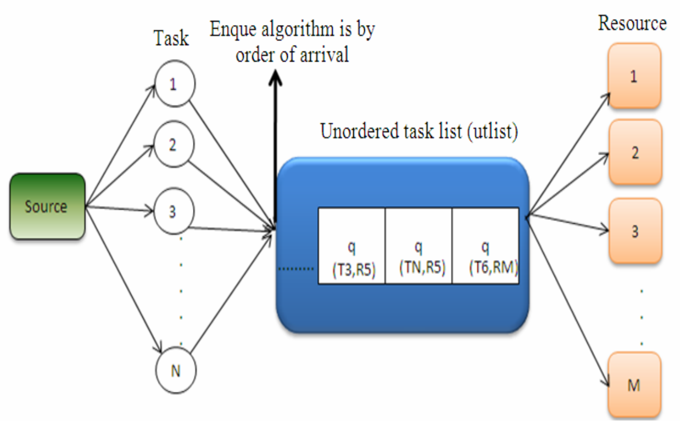

Task
Enque algorithm is by order of arrival
Resource
Source
1
2
3
N
Unordered task list (utlist)
q (T3,R5)
q (TN,R5)
q (T6,RM)
1
2
3
M

Fig. 4. Interaction of Entities and Resources

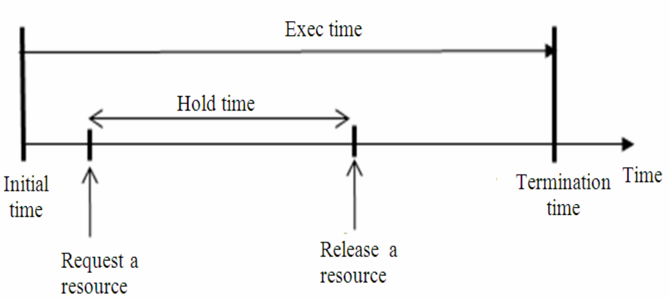


Fig. 5. Task Model

During the lifetime of a task, it may request one or more resources. For each request, a task specifies the duration to hold the requested resource. This is denoted as Hold time. The Exec time denotes the remaining execution time of a task at a particular instant. Initially, at Initial time the value of Exec time is equal to C_AVG. This value is reduced as the task is executed until the Termination time and the value of Exec time becomes zero. It is assumed that a task releases all resources it acquires before it ends, complying with condition of the Hold time ≤ Exec time. The following assumptions are made for the task model implemented in this research:

- Independent task model, whereas each task has no dependency on other task during execution. The execution of a task has no correlation to the previously executed task.
- Task can be preemptive, i.e., a task can be delayed or suspended to allow another task to be executed.

2.1 TUF Model

The timing constraint of a task is designed using the step TUF model in this research [4]. A TUF describes a task contribution to the system as a function of its completion time. The step TUF model is shown in Fig. 1. The maximum utility that could possibly be gained by a task is denoted as MaxAU. The random value of MaxAU abides normal distribution (10, 10) i.e., the mean value and variance is set 10 to conform to the benchmark. The Initial time is the starting time for which the function is defined. The Termination time is the latest time for which the function is defined. That is, MaxAU is defined in within the time interval of [Initial time, Termination time]. The completion of a task within this interval will yield positive utility i.e., MaxAU to the system. The completion of a task breaching the stipulated deadline causes the value of MaxAU to become zero. If the Termination time is reached and the task has not finished its execution, it accrues zero utility to the system.

The constant amount of resources and surplusing demands results in resource unavailability. The simulator provides a mechanism to retain the task's requests for resources which are temporarily unavailable in an unordered task list named as utlist. A queue implementation via the pointer based single link list is used to deploy the utlist as shown in Fig. 6.

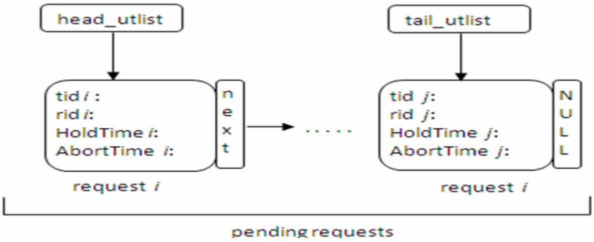


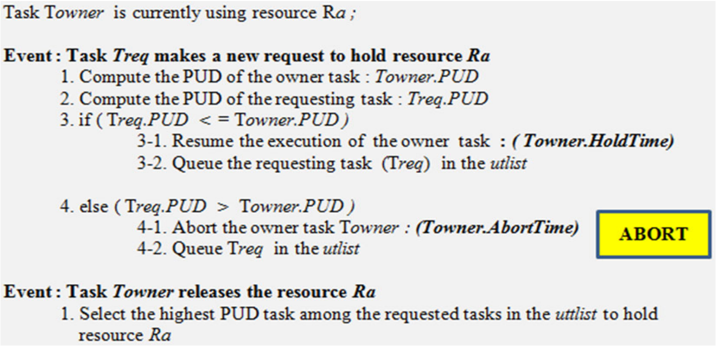
Fig. 6. Queuing Model

Referring to Fig. 6, the utlist consists of a sequence of pending request. A request for a resource is represented by a quadruple ReqResourceItem=<tid,rid, Holdtime, AbortTime>. Thus, an element in the utlist consists of ReqResourceItem structure. A next pointer is used to link an element to the next element in the utlist. The head_utlist points to the first element and tail_utlist points to the final element in the utlist.

2.2 Scheduling Algorithm Component

The scheduling algorithms component consists of the benchmark GUS algorithm. GUS is a TUF/UA uniprocessor scheduling algorithm that considers the step and arbitrary shape TUFs. The main objective of GUS is to maximize the utility accrued to represents that the most important task is to be scheduled first in the system. GUS uses a greedy strategy where task whose execution yields the maximum PUD over others is selected to determine which task to be scheduled at a particular instant.

The PUD of a task measures the amount of utility that can be accrued per unit time by executing the task. It essentially measures the Return on Investment (RoI) for executing the task at current clock time. Fig. 7 elaborates the GUS scheduling algorithm for the execution of an independent task model.



```
Task Towner is currently using resource Ra ;

Event : Task Treq makes a new request to hold resource Ra
        1. Compute the PUD of the owner task : Towner.PUD
        2. Compute the PUD of the requesting task : Treq.PUD
        3. if ( Treq.PUD < = Towner.PUD )
                3-1. Resume the execution of the owner task : ( Towner.HoldTime)
                3-2. Queue the requesting task (Treq) in the utlist

        4. else ( Treq.PUD > Towner.PUD )
                4-1. Abort the owner task Towner : (Towner.AbortTime)
                4-2. Queue Treq in the utlist

Event : Task Towner releases the resource Ra
        1. Select the highest PUD task among the requested tasks in the uttlist to hold
           resource Ra
```

Fig. 7. GUS Scheduling Algorithm [3]

When a new request from task Treq arrives into the system, GUS accepts the new request for resource Ra. Referring to Fig. 7, when the resource Ra is currently being used by another task i.e., task Towner, GUS firstly calculates the PUD of both tasks. In the case that the requesting task i.e., Treq posses a higher PUD as compared to task Towner, GUS has tailored mechanism to abort the current owner task (i.e., Towner) and grant the resource to the requesting task (i.e., Treq).

3 Experimental Settings

Extensive experiments were done to ensure the developed GUS simulator is validated. The simulation model is validated by ensuring that its output data closely resemble the output data that was observed from the benchmark model i.e., GUS. The developed simulator has been tailored to map the characteristics of a uniprocessor scheduling. Table 1 summarizes the simulation parameter settings that are used throughout this research [4].

Table 1. Experimental Settings

Parameters	Value
load	0.20 to 1.60
iat	Exponential (C_AVG/load)
Hold time	Normal (0.25,0.25)
MaxAU	Normal(10,10)
Abort time	Any random number < Hold time

A source generates a stream of 1000 tasks. Given the task average execution time C_AVG and a load factor load, the average task inter arrival time i.e., iat is calculated as the division of C_AVG over load and further utilized an exponential distribution to be further derived to reflect the intended system model. In all the simulation experiments, the value of C_AVG is set at 0.50 sec and the range value of load is from 0.20-1.60. The different values of load are to provide the derivation of differing mean arrival rates of tasks. The arrival of tasks is assumed to follow the exponential distribution. The system is said to be overloaded when (load >1.00) represented also as the mean arrival rate of 0.50 seconds (i.e., iat). This complementary representation of load can be utilized to show congestion as the iat is at its equal value to the execution ability to process a task. The value of the HoldTime and AbortTime parameters are derived by the normal distribution with mean and variance is 0.25. The maximum utility of a task i.e., MaxAU is computed using normal distribution with mean value of 10 and variance of 10.

The performances of real time scheduling algorithms are measured by the metrics which rely on the respective application specifications. The Accrued Utility Ratio (AUR) metric defined in [1] has been extensively utilized in the existing TUF/UA scheduling algorithms and is considered as the standard metric in this domain [2],[3],[4]. AUR is defined as the ratio of accrued aggregate utility to the maximum possibly attained utility.

4 Result and Conclusion

A result obtained from simulator is compared with the result published in the literature by using the same assumptions and experimental settings [4].

Fig.8 depicts the AUR results of the developed GUS simulator and the original GUS. The result obtained using the simulation is comparable to the result published with the same trends. From the results, as the number of load is increased; a lower accrued utility is recorded. In addition, the respective results and the deviation between the developed GUS simulator and the benchmark model is validated with less than 5% as compared to the benchmark model.

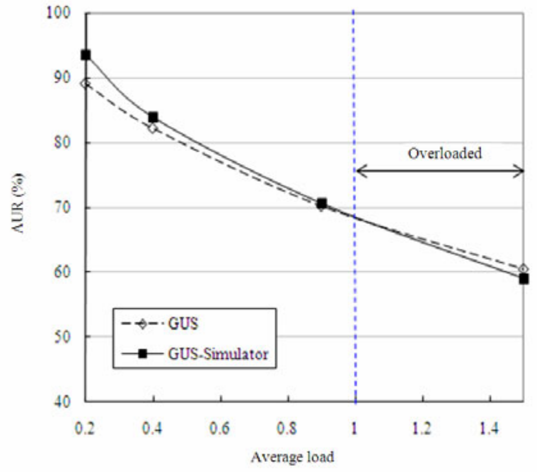


Fig. 8. AUR Result of the developed simulator and the benchmark model

This study has provided the design of the developed GUS scheduling algorithm by using DES. The aim of the developed simulation framework was not only to develop a GUS model for the research problem but also to provide a platform for future investigations involving TUF/UA real time scheduling. For future work, the GUS algorithm can be deployed in network and distributed environment. Flow control and routing algorithms should be integrated into the model to increase the feasibility in actual implementation of the algorithm.

References

1. Jensen, D., Locke, D., Tokuda, H.: A Time Driven Scheduling Model for Real Time Operating Systems. In: IEEE Symposium on Real Time System, pp. 112–122. IEEE Press, New York (1985)
2. Wu, H., Ravindran, B., Jensen, D., Li, P.: CPU Scheduling for Statistically Assured Real Time Performance and Improved Energy Efficiency. In: 2nd IEEE/ACM/ IFIP International Conference on Hardware/Software Codesign and System Synthesis, pp. 110–115. IEEE Press, New York (2004)
3. Li, P., Wu, H., Ravindran, B., Jensen, D.: A Utility Accrual Scheduling Algorithm for Real- Time Activities with Mutual Exclusion Resource Constraints. IEEE Trans. Computer 55, 454–469 (2006)
4. Law, A.: How to Conduct a Successful Simulation Study. In: Winter Simulation Conference, pp. 66–77 (2003)

Engineering Design Analysis Using Evolutionary Grammars with Kano's Model to Refine Product Design Strategies

Ho Cheong Lee

School of Computer Systems and Software Engineering,
University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang,
Kuantan, Pahang Darul Makmur, Malaysia
jackielee@ump.edu.my, jackielee2005@gmail.com

Abstract. The ability for a product developer to successfully launch useful products to a market is tied to the company's product development strategies, thus making profitability. This paper investigates the shape formulation process for product design strategies. The research focuses on nonlinear product design analysis with Kano's model to refine product design strategies using interactive evolutionary grammars based design framework.

In analyzing the generated designs, Kano's attribute curves (Basic, Performance and Exciting) are mapped to the analysis results to form reference models. By comparison of the user preference curves with the Kano's reference models, patterns emerged to delineate different approaches to the product design strategies like fostering innovation and creativity in product design, controlling production expense, searching the targeted users, enhancing or lowering functionalities, and increasing or decreasing resources, features and services. Upon determining the appropriate Kano's reference models, product developers could refine product design strategies to suit the targeted market.

Keywords: Evolutionary shape grammars; Genetic programming; Product design; Engineering design; Nonlinear product design analysis; Kano's model; Product development strategies; Customer satisfaction.

1 Introduction

Shape formulation is a critical issue in Engineering Design. Over thirty years' research on shape grammars has established a solid theoretic foundation in shape formulation for various domains like architecture, structural and engineering design. A comprehensive survey which compared the development processes, application areas and interaction features of different shape grammar approaches is given by Chase (2002) [1]. Recently, research in exploring shape grammar approach to product and engineering design has received more and more attention by many researchers. For instance, Cagan et al. developed the coffeemaker grammar, motorcycle grammar, hood panel grammar and vehicle grammar [2], [3], [4] and [5].

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 627–641, 2011.

Recently, Lee et al. (2008) researched on the advanced evolutionary framework which utilizes the power of evolutionary algorithms and shape grammars in generating innovative designs [6]. The evaluation criteria could be so complex with many multi-dimensional variables. This leads the product developers to have difficulties in making decisions by interpreting the analysis results.

In order to tackle this problem, this research realizes the implications on the analysis results by comparing them against the attribute curves of Kano's model. The Kano's model is developed in the 80s by Professor Noriaki Kano to define product development strategies in relation to customer satisfaction which classifies customer preferences into five categories: Attractive, One-Dimensional, Must-Be, Indifferent and Reverse [7]. The reasons for mapping Kano's attribute curves to the product design analysis results are: 1) to redevelop or modify the control strategies of the framework by the shape grammar developer, and 2) to refine or adjust the product design strategies by the product developer.

Section 2 reviews related research works. Section 3 presents the development of the framework with the Kano's reference models. Section 4 illustrates the implementation of this framework and analyses the results. Finally, section 5 draws conclusions.

2 Related Work

The recent related research on applying evolutionary algorithms and shape grammars to engineering design are reviewed and divided into four issues: 1) Planning, 2) Evolution, 3) Control, and 4) Evaluation.

Planning is the starting design activity in which theoretical design concepts by means of research are derived and practical skills of experts are quantified for computation. The quantified objects in terms of parameters, variables and transformation rules should be defined in order to build up the language of shape grammars for design applications. The language of shape grammar consists of a vocabulary of shape elements, a set of production (or transition) rules and an initial shape. For instance, Hohmann et al. (2010) formulated shape grammars using the Generative Modeling Language (GML) from Havemann (2005) to build semantically enriched 3D building models for facility surveillance application [8] and [9].

Evolution is the design activity in which a blueprint of exploration is established and implemented in a well controlled environment. Recent research on evolutionary design is focused on adding generative capability to existing Computer-Aided Design (CAD) systems. For instance, Krish (2011) demonstrated the Generative Design Method (GDM) based on parametric search and evolutionary algorithms to create unanticipated solutions in history based parametric CAD systems [10].

Control is the design activity in which the rate of exploration determines the contraction or expansion of the design space. The design space can be contracted to avoid a radical change in generating the designs, or expanded for generating new designs without any constraints. Apart from the emergent property of shape grammars described in Stiny's recent book [11], another advantage of applying shape grammar is to control the design process. As design problems are "ill-defined" as determined by Simon (1984, 1990) [12] and [13], control of design process is a critical issue in engineering design. Dorst (2006) further emphasized the issues of specifying appropriate design problems with right kind of abstractions and correctness [14].

Recent research on shape grammars in CAD, Computer-Aided Process Planning (CAPP) and Computer-Aided Manufacturing (CAM) are focused on the integration of shape grammars and cognitive approaches in automation, and utilization of emergent properties of shape grammars. For instance, Shea et al. (2010) has developed new fabrication systems with flexible and cognitive capabilities for autonomous design-to-fabrication automation [15]. Fox (2010) has proposed the development of Unified Shape Production Languages with the power of emergent properties of shape grammars for sustainable product creation [16]. However, these approaches do not much address the controllability issue in shape formulation. Lee et al. (2008) has developed control strategies to tackle this issue by interactively determining the exploration rate to modify the shape grammar rules [6].

Evaluation is the design activity in which the results are evaluated by a set of evaluation criteria. The Kano's model is adopted in this research to enhance the analytical ability of evolutionary grammars. The Kano's model defines product development strategies in relation to customer satisfaction and expectations on product requirements. The expectations could include categorizations of requirements that relate to technical management [17]. For instance, the common categorizations of requirements include Customer, Architectural, Structural, Behavioral, Functional, Non-functional, Performance, Design, Derived and Allocated requirements.

Recent research on the Kano's model covers a vast variety of applications. For instance, Alessandro et al. (2009) has identified a nonlinear and asymmetric relationship between attribute performances and overall customer satisfaction using a case study of retail banking [18]. Cheng et al. (2009) has applied the Kano's method to extracting the implicit needs from users in different clusters which were grouped by the artificial neural networks [19]. This approach was applied to content recommendation in web personalization. The results indicated that the problem of information overloading was improved.

3 Development of the Framework with the Kano's Reference Models

The development of the framework is first introduced with illustration by a case study using digital camera form design. It follows with the methodologies in mapping the Kano's attribute curves to the performance graph to form Kano's reference models. The reference models could then be used to refine product design strategies.

3.1 Interactive Evolutionary Grammar Based Design Framework

Parametric 3D shape grammars with labels are developed which follow the generative specification of the class of compact digital camera forms as illustrated in Figure 1.

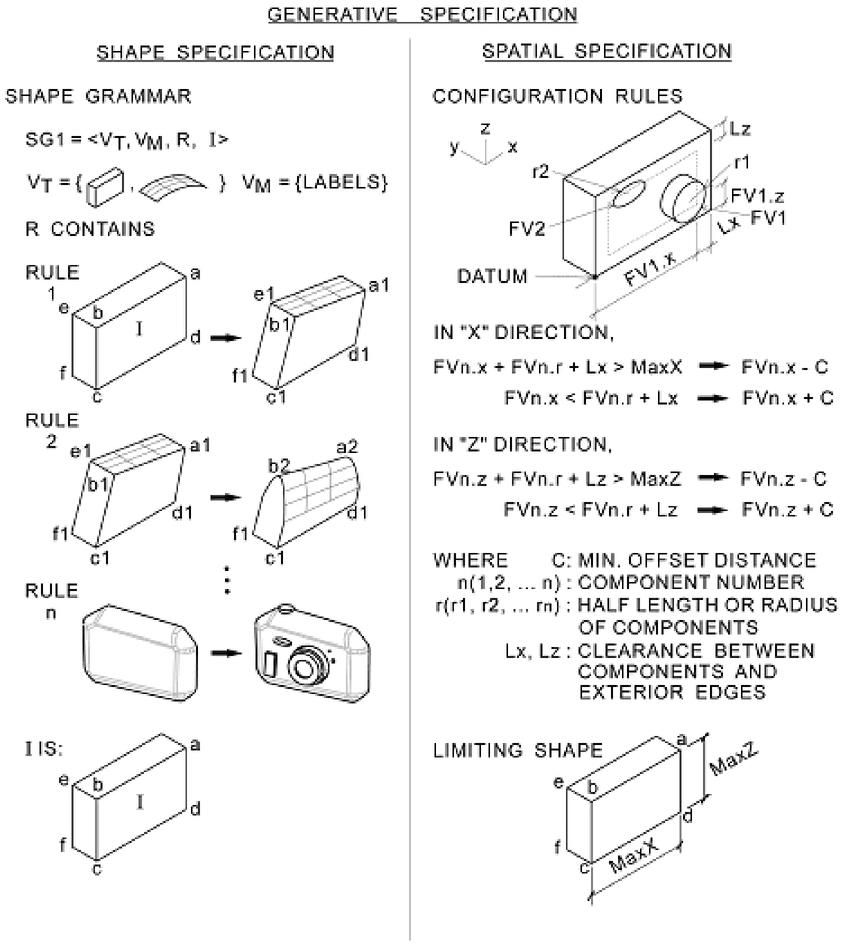


Fig. 1. Generative specification of the class of compact digital camera forms

Figure 2 shows the framework in which genetic programming is selected to explore and optimize product form designs. Control strategies are also developed in manipulating the genetic representation and systematically evaluating the evolving designs during the evolutionary process.

Exterior form generation of compact digital cameras and the configuration of the components are designed to fulfill a set of requirements such as artificial selection, spatial geometric constraints and desired exterior shell volume. The design requirements can be formulated into objective functions. Objective functions are set up for the evaluation of the generated designs.

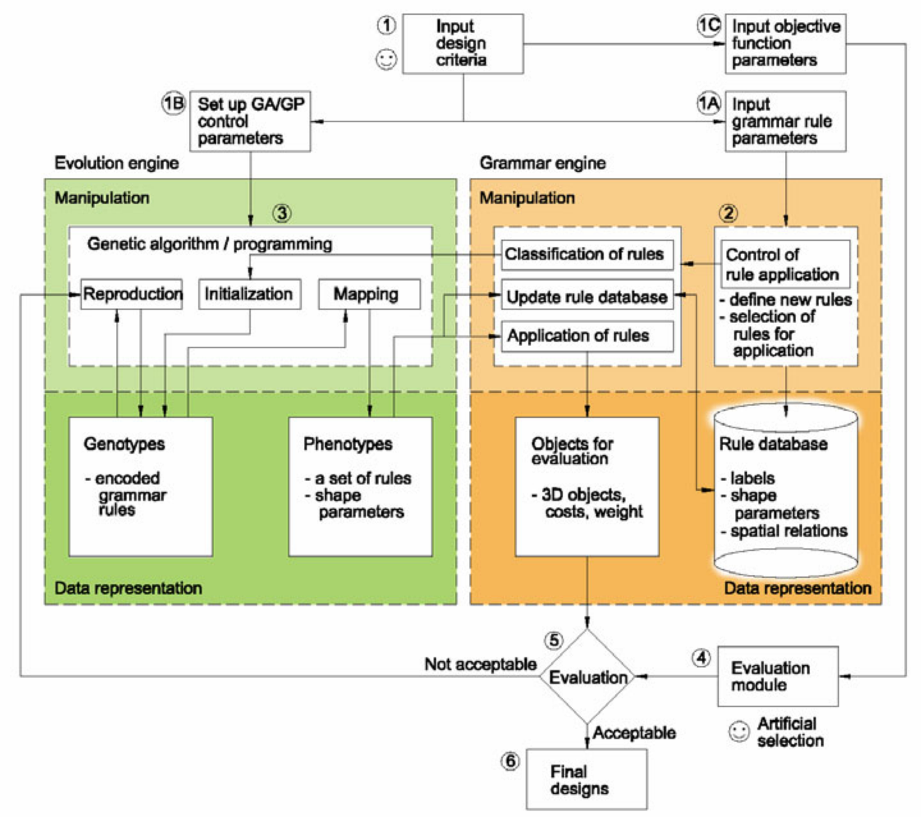


Fig. 2. Evolutionary grammar based design framework

General objective functions are set up for general requirements while control strategies have their own sets of objective functions for specific requirements. Analysis of the evaluation results will help in the investigation of and understanding of combinatorial effects on the generated designs based on the control strategies. To effectively evaluate the design performance, a metric is formulated as the summation of design objectives and weighted constraint violations.

$$\begin{aligned}\text{Index function} &= \text{Objective index} + \text{Constraint index}\\ &= \sum_{i=1}^{l} \text{Objective index}_i + \sum_{j=1}^{m} \text{Constraint index}_j \qquad (1)\end{aligned}$$

where : l = number of objectives,

m = number of constraints.

Objective and penalty functions are defined to assign positive and negative fitness scores respectively. Penalty functions are activated if the generated designs violate the constraints. Both design objectives and constraints have weighting factors to determine the relative trade-off among design objectives. The designers can assign different weighting factors on each variable.

$$\text{Objective index} = \sum_{i=1}^{l}(\text{Objective weight}_i \bullet \text{Objective value}_i) \tag{2}$$

where : l = number of objectives.

$$\text{Constraint index} = \sum_{j=1}^{m}(\text{Constraint weight}_j \bullet \text{Constraint violation}_j) \tag{3}$$

where : m = number of constraints.

For the artificial selection requirements, *Objective index*$_1$ is used as the measurement of accumulated effect on selected designs. The selected designs will be assigned with higher fitness scores if they are frequently selected by the designers.

$$\text{Objective index}_1 = \sum_{i=1}^{n}(\text{Selection weight}_i \bullet \text{Selection value}_i)$$
$$\{\text{Selection value}_i = 0 \text{ or } 1\} \tag{4}$$

where n = number of generations; *Objective index*$_1$ is the accumulated score for each design; *Selection weight*$_i$ is the weighting factor for each design; *Selection value*$_i$ is assigned with 1 when the designs are selected, otherwise 0. Since the selection cost of each design is the accumulated score from each generation, selection on one or more designs in a particular generation will not significantly impact the whole population. As a result, the population is determined by the accumulated effect on the selected designs.

Under the spatial geometric constraints, the components have to be configured without collision among each other and within the boundary of the exterior of camera body. Geometric variables of the component positions and the boundary positions of the exterior of the camera body are assumed to be configuration design variables, subject to a set of constraints. The objective functions of configuration of components can be determined by the designers with selective options. For example, the selective options of configuration are: to maximize or minimize the total distance (TD_1) among components.

For Maximize option selected :

$$\text{Objective index}_2 = \text{Configuration weight} \bullet \text{TD}_1 \tag{5}$$

For Minimize option selected :

$$\text{Objective index}_2 = \text{Configuration weight} \bullet \frac{1}{\text{TD}_1 + \text{C}} \tag{6}$$

$$\begin{aligned}&\mathrm{TD}_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij},\ \{i \neq j\}\\ &\text{Subject to (a set of constraints)}:\\ &d_{ij} \geq l_i + l_j + l_c,\ \{i = 1 \text{ or } 2 \text{ or },\ldots, \text{ or } n\},\\ &\{j = 1 \text{ or } 2 \text{ or },\ldots, \text{ or } n\} \text{ and } \{i \neq j\}\end{aligned} \tag{7}$$

$$\begin{aligned}&\text{Constraint index}_1\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\text{Configuration constraint weight} \bullet \text{Constraint violation}_{ij})\\ &\{\text{Constraint violation}_{ij} = -(l_i + l_j + l_c - d_{ij}), \text{ if the constraints are vi}\\ &\{\text{Constraint violation}_{ij} = 0, \text{ if the constraints are not violated}\}.\end{aligned} \tag{8}$$

where C is a constant; n is the number of components; l_i , l_j are the half length or radius of components; l_c is the clearance between components; coefficient d_{ij} is the distance between components i and j . The distance between two components is defined as the distance between the centres of both components. The summation of all the distances between any two components (TD_1) reflects the dispersion among components.

For exterior shell volume calculation, the objective is to minimize the difference between the shell volume and a desired target shell volume of the exterior of camera body.

$$\text{Objective index}_3 = (\text{Volume weight} \bullet \mathrm{f(v)}) \tag{9}$$

$$\text{Minimise } \mathrm{f(v)} = \frac{1}{\left|(\mathrm{v} - \mathrm{v}_{\mathrm{target}})\right| + \mathrm{C}} \tag{10}$$

where : C = constant.

The value of an exterior shell, v , refers to the approximate volume estimation of the exterior of the camera body. A constant C is added to $Objective\ index_2$ and $f(v)$ to ensure that the objective indices take only positive values in their domains [20]. The addition of constant C to the objective indices also avoids the error arising from dividing zero.

3.2 Mapping Kano's Attribute Curves to the Nonlinear Time Dependent Product Design Analysis Results to form Kano's Reference Models

The Kano's model provides the determination of importance on the product attributes from the customer's point of view. This is achieved by quantifying the relationship between the quality performance of a product or service (X-axis in Figure 3) and customer satisfaction (Y-axis in Figure 3). The Kano's model allows the product development team to better understand and discuss the issues in defining the specifications of the design problems. Figure 3 depicts three kinds of time independent attribute curves which represent conceptual relationships including Basic, Performance and Exciting relationships.

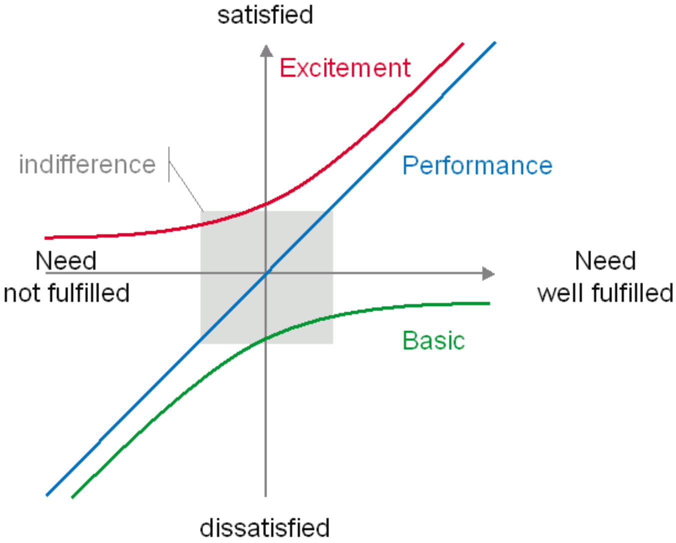


Fig. 3. Kano's model

On the other hand, the product designs are generated and analyzed in nonlinear time dependent manner by the evolutionary framework. The analyzed results are time dependent and are plotted in the performance graph. This induces a technical problem that the Kano's attribute curves could not directly map to the performance graph. Therefore, assumptions are made on the Kano's attribute curves to suit the purpose in this application.

Refer to Figure 3, the Basic attribute curve refers to an essential performance of the product (X-axis) which is sufficient to satisfy the customers' need (Y-axis). It is assumed that the performance of the product (X-axis) is gradually improved along time using the same scale of generation. For instance, the performance of the product (X-axis) is started with zero at the beginning of evolutionary process (0th generation) and ended with 500 at the end of the evolutionary process (500th generation). Similarly, the Performance attribute curve demonstrates a linear relationship in improving the satisfaction (Y-axis) with respect to performance of the product or time (X-axis). For the case of Exciting attribute curve, increasing performance of the product during the evolutionary process (X-axis) can produce more customer satisfaction (Y-axis). These three Kano's attribute curves can be mapped to the performance graph to form Kano's reference models respectively (Figure 4 to 6).

3.3 Comparison Analysis of Kano's Reference Models

After forming the Kano's reference models, four typical cases of user preferences are proposed and simulated for testing the reference models. The user preference curves are simulated by modifying the Artificial Selection Fitness parameters of the performance graph. Finally, comparisons analysis on the User overall fitness curves (User-1, User-2, User-3 and User-4) against the three Kano's reference models could be made respectively (Figure 4 to 6). The key operation procedures of the framework are summarized as follow:

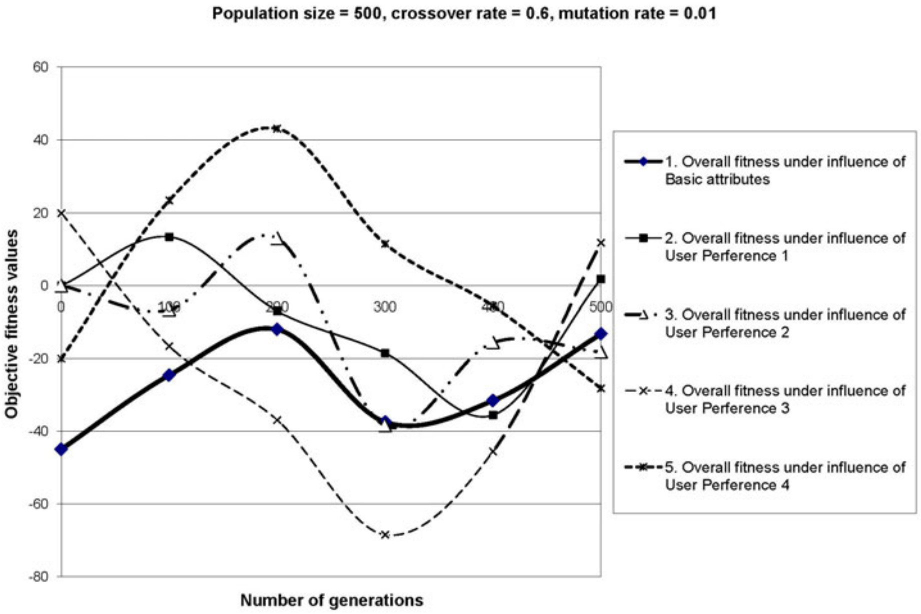


Fig. 4. Comparison of Kano's Basic reference model

Step 1) Generate the product designs and analyzed by the evolutionary framework;

Step 2) Plot the analyzed results in the performance graph;

Step 3) Map the three Kano's attribute curves: Basic, Performance and Exciting attribute curves to the performance graph to form reference models;

Step 4) Modify the Artificial Selection Fitness parameters of the performance graph to simulate four overall fitness curves under the influence of four user groups;

Step 5) Compare the user overall fitness curves (User-1, 2, 3 and 4 preferences) against the three Kano's reference models respectively (Figure 4 to 6); and

Step 6) Decision making upon the comparison analysis of the Kano's reference models (see Section 4, Table 1).

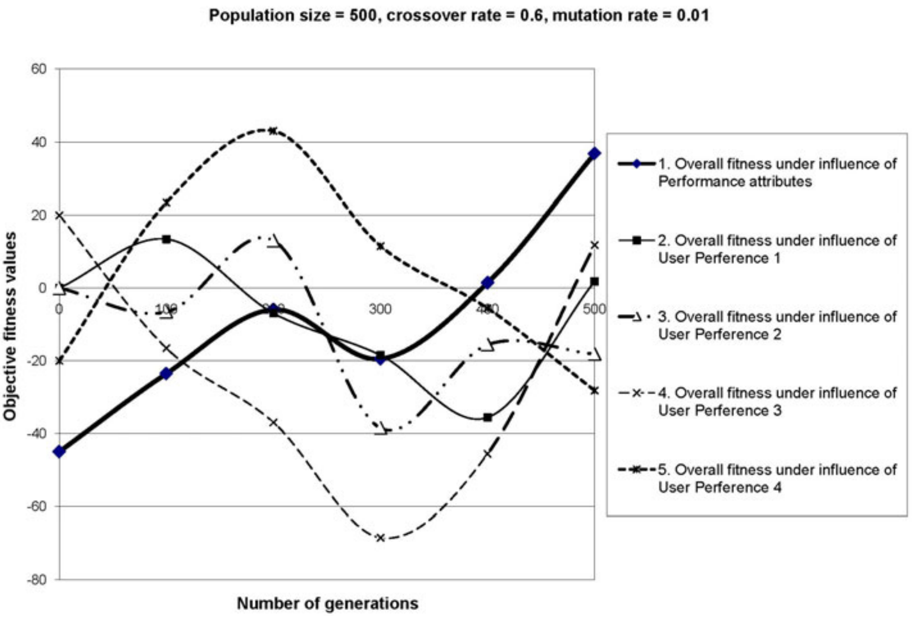


Fig. 5. Comparison of Kano's Performance reference model

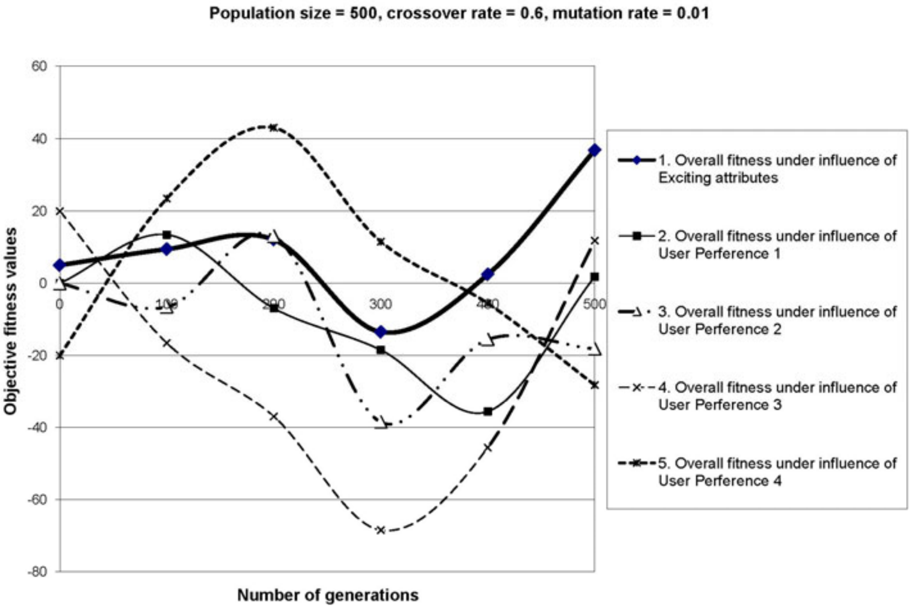


Fig. 6. Comparison of Kano's Exciting reference model

4 Implementation

A software prototype system has been developed using Visual C++ and ACIS 3D modelling kernel, and tested. Figure 7 shows the implementation results obtained from the system, starting at the first generation and ending at five hundred generations.

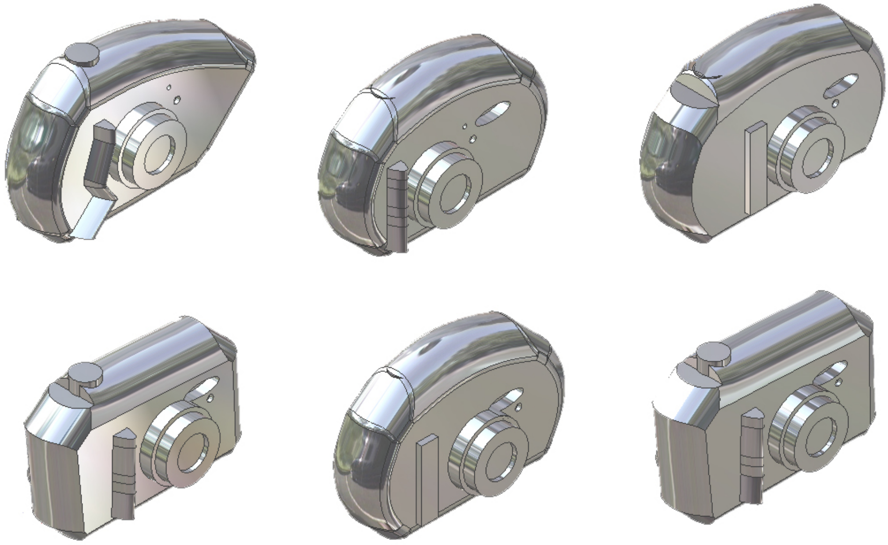


Fig. 7. Results obtained from the first generation (top left), 100 generations (top middle), 200 generations (top right), 300 generations (bottom left), 400 generations (bottom middle) and 500 generations (bottom right)

4.1 Interpretation of Kano's Reference Models to Refine Product Development Strategies

Decisions on refinement of the product design strategies could be made by the product developers upon the completion of comparison analysis of the Kano's reference models (Table 1). The selection of which Kano's reference model for comparison is depended on the product developer strategies. For instance, targeting to sell products to the general public in underdeveloped countries, the expected product cost must be comparatively lower than to developed countries. As a result, the Kano's Basic reference model could probably be adopted by the developer.

4.2 Comparison with Basic Reference Model

As depicted in Figure 4, User-1, User-2 and User-4 overall fitness curves are over Kano's Basic reference model. This means that the generated designs outperform the essential performance of the product (X-axis) which leads to more customer satisfaction (Y-axis) than the expectation. Profitability would deteriorate somewhat when launching these new products in future, largely due to an increase in production expense percentage. The reasons are that the excess resources, services or functions are provided to the generated designs which cause higher production costs. Solutions may then be proposed by the product developer to refine the product design strategies. For instance, reducing the excess functions on the generated designs would lower down the essential performance of the product (X-axis). The expenses on production would then be reduced and the profitability be improved.

Figure 4 also reveals that User-3 overall fitness curve is under Kano's Basic reference model from generation 100 to generation 400. This indicates that the generated designs underperform the essential performance of the product (X-axis) which leads to less customer satisfaction (Y-axis) than the expectation. Profitability would depreciate somewhat when launching these new products to the market, largely caused by a decline in sales volume. The causes behind that are needed to be identified from User-3 preference curve. For instance, the customer satisfaction (Y-axis) of User-3 preference curve is lower at generation 200 to generation 300. This implies that lack of resources, services or functions are provided to the generated designs at this generation period which would cause a decline in sales volume. The product developer may request the shape grammar developer to make minor modification on the control strategies of the framework. For instance, major improvements on the customer satisfaction (Y-axis) could then be obtained at generation 200 to generation 300 by raising the functions on the generated designs at that generation period. The expenses on production could be higher but it would be compensated by the enhancement of the profitability.

4.3 Comparison with Performance Reference Model

There are no obvious relationships to be found among User-2, User-3 and User-4 overall fitness curves and Kano's Performance reference model as depicted in Figure 5. The product developer may need to refine the product design strategies for each of these User overall fitness curves. Except that for User-1 overall fitness curve, the closest match to the reference model appears at generation 200 to generation 300. The features and functions produced as well as the User-1 categories at this generation period may be focused by the product developer. A linear relationship between the reasonable production expenses and the profitability could then be obtained.

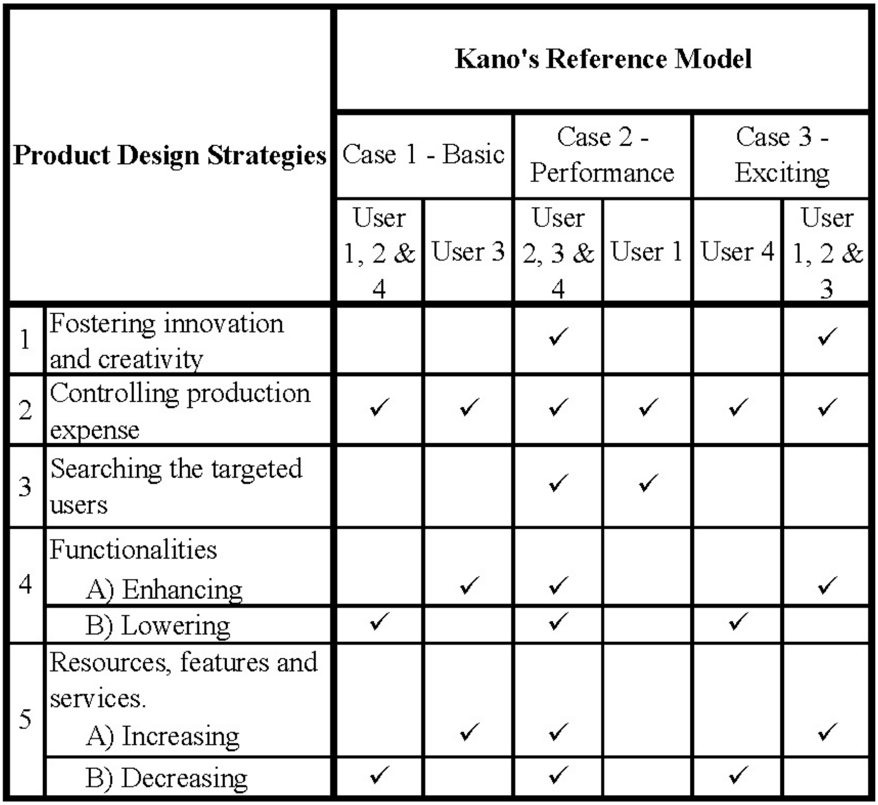
4.4 Comparison with Exciting Reference Model

As depicted in Figure 6, User-4 overall fitness curve is over Kano's Exciting reference model from generation 100 to generation 400 whereas User-1, User-2 and User-3 are under. Again, identification of the causes behind that is necessary for refinement of product strategies. For instance, the customer satisfaction (Y-axis) of User-4 preference curve is higher at generation 200 to generation 300. This implies that excess resources, services or functions are provided to the generated designs at this generation period which would cause an increase in production expense percentage. The product developer may propose to reduce the excess functions on the generated designs at that generation period for User-4 overall fitness curve. This would lower down the essential performance of the product (X-axis). The expenses on production would then be reduced and the profitability be improved.

For the cases of User-1, User-2 and User-3 overall fitness curves, major improvements on the customer satisfaction (Y-axis) could then be obtained by raising the functions on the generated designs. The expenses on production could be higher but it would be compensated by the enhancement of the profitability.

A summary of product development strategies in this case study is shown in Table 1.

Table 1. Decisions on refinement of the product design strategies



Product Design Strategies		**Kano's Reference Model**					
		Case 1 - Basic		Case 2 - Performance		Case 3 - Exciting	
		User 1, 2 & 4	User 3	User 2, 3 & 4	User 1	User 4	User 1, 2 & 3
1	Fostering innovation and creativity			✓			✓
2	Controlling production expense	✓	✓	✓	✓	✓	✓
3	Searching the targeted users			✓	✓		
4	Functionalities A) Enhancing		✓	✓			✓
	B) Lowering	✓		✓		✓	
5	Resources, features and services. A) Increasing		✓	✓			✓
	B) Decreasing	✓		✓		✓	

5 Conclusion

Innovation and creativity are the key successful factors and a global priority in engineering industries. One of the issues in generating innovative and creative designs is to define appropriate evaluation criteria. The evaluation criteria could be so complex with many multi-dimensional variables. This leads the product developers to have difficulties in making decisions by interpreting the analysis results. This paper addresses this issue with a new framework which incorporates with non linear product design analysis and Kano's reference models for engineering design applications.

For Kano's Basic reference model, excess features are not necessary provided to the products in order to maintain reasonable production expense. For instance, features adopted with high technology are usually not preferred to provide to the product. Instead, providing a variety of innovative stylistic designs with acceptable quality would be chosen.

For Kano's Performance reference model, providing excess features and better functions to the products would result in customer satisfaction. Conversely, providing lack of features and poor functions to the products reduces customer satisfaction. Risk management on an increase of production expense in providing excess features and better functions to the product should be considered. For instance, would an increase of the price for the product for the excess features and better functions deter customers from purchasing it?

For Kano's Exciting reference model, excess features may be necessary provided to the products in order to unexpectedly delight customers. For instance, features adopted with high technology are generally preferred to provide to the product. In addition, providing alternative innovative stylistic designs with excellent quality would also be chosen. Examples of this category of products are iPad, iPhone, Tablet PC and Portable game console with an autostereoscopic three-dimensional effect (one without 3D glasses). Somehow, risk management techniques should also be considered in using Kano's Exciting reference model. For instance, in case most users do not accept the product with the features adopted with new technology, promotion of such features should be carried out. It is a difficult task to advertise new features to the customers without clear explanation on the advantages of those features. As a result, costs of manpower in promotion activities should be taken into account in applying the Kano's Exciting reference model.

Product development strategies have been provided in this research upon decision making based on the interpretation of the Kano's reference models. At this stage, the results are analyzed artificially based on visual interpretation by comparing the significant differences among the User overall fitness curves against the three Kano's reference models respectively. Further research will be considered in enhancing the performance of the framework by implicitly embedding the Kano's model within the framework rather than explicitly map the Kano's functions onto the performance graph. This allows the analysis of the generated products more precisely to reflect the dynamic change on user's satisfaction.

References

1. Chase, S.C.: A Model for User Interaction in Grammar-based Design Systems. Automation in Construction 11, 161–172 (2002)
2. Agarwal, M., Cagan, J.: A Blend of Different Tastes: The Language of Coffeemakers. Environment and Planning B: Planning and Design 25, 205–226 (1998)
3. Pugliese, M.J., Cagan, J.: Capturing a Rebel: Modelling the Harley-Davidson Brand through a Motorcycle Shape Grammar. Research in Engineering Design: Theory Applications and Concurrent Engineering 13(3), 139–156 (2002)

4. McCormack, J.P., Cagan, J.: Designing Inner Hood Panels through a Shape Grammar Based Framework. AI EDAM 16(4), 273–290 (2002)
5. Orsborn, S., Cagan, J., Pawlicki, R., Smith, R.C.: Creating Cross-over Vehicles: Defining and Combining Vehicle Casses using Shape Grammars. In: AIEDAM, vol. 20, pp. 217–246 (2006)
6. Lee, H.C., Tang, M.X.: Evolving Product Form Designs using Parametric Shape Grammars Integrated with Genetic Programming. In: Artificial Intelligence for Engineering Design, Analysis and Manufacturing, vol. 23, pp. 131–158. Cambridge University Press, Cambridge (2008)
7. Kano, N., Seraku, N., Takahaashi, F., Tsuji, S.: Attractive Quality and Must-be Quality, Hinshitsu. The Journal of the Japanese Society for Quality Control 14(2), 39–48 (1984)
8. Hohmann, B., Havemann, S., Krispel, U., Fellner, D.: A GML Shape Grammar for Semantically Enriched 3D Building Models, Computers & Graphics. Procedural Methods in Computer Graphics; Illustrative Visualization 34(4), 322–334 (2010)
9. Havemann, S.: Generative Mesh Modeling. PhD thesis, Technical University Braunschweig
10. Krish, S.: A Practical Generative Design Method. Computer-Aided Design 43(1), 88–100 (2011)
11. Stiny, G.: Shape, Talking about Seeing and Doing. The MIT Press, Cambridge (2006)
12. Simon, H.A.: The Structure of Ill-structured Problems. In: Nigel, C. (ed.) Developments in Design Methodology, John Wiley & Sons, New York (1984)
13. Simon, H.A.: The Sciences of the Artificial, 2nd edn. The MIT Press, Cambridge (1990)
14. Dorst, K.: Design Problems and Design Paradoxes. Design Issues 22(3), 4–17 (2006)
15. Shea, K., Ertelt, C., Gmeiner, T., Ameri, F.: Design-to-Fabrication Automation for the Cognitive Machine Shop, Advanced Engineering Informatics. The Cognitive Factory 24(3), 251–268 (2010)
16. Fox, S.: Generative Production Systems for Sustainable Product Creation. Technical Research Centre of Finland, VTT (2009)
17. Systems Engineering Fundamentals. Defense Acquisition University Press (2001)
18. Arbore, A., Busacca, B.: Customer Satisfaction and Dissatisfaction in Retail Banking: Exploring the Asymmetric Impact of Attribute Performances. Journal of Retailing and Consumer Services 16(4), 271–280 (2009)
19. Chang, C.C., Chen, P.L., Chiu, F.R., Chen, Y.K.: Application of Neural Networks and Kano's Method to Content Recommendation in Web Personalization. Expert Systems with Applications 36(3) Part 1, 5310–5316 (2009)
20. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, London (1996)

Applications of Interval-Valued Intuitionistic Fuzzy Soft Sets in a Decision Making Problem

Xiuqin Ma, Norrozila Sulaiman, and Mamta Rani

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
xueener@gmail.com, norrozila@ump.edu.my, mamta@ump.edu.my

Abstract. Soft set theory in combination with the interval-valued intuitionistic fuzzy set has been proposed as the concept of the interval-valued intuitionistic fuzzy soft set. However, up to the present, few documents have focused on practical applications of the interval-valued fuzzy intuitionistic soft sets. In this paper, Firstly, we present the algorithm to solve decision making problems based on interval-valued intuitionistic fuzzy soft sets, which can help decision maker obtain the optical choice. And then we propose a definition of normal parameter reduction of interval-valued intuitionistic fuzzy soft sets, which considers the problems of suboptimal choice and added parameter set, and give a heuristic algorithm to achieve the normal parameter reduction of interval-valued intuitionistic fuzzy soft sets. Finally, an illustrative example is employed to show our contribution.

Keywords: Soft sets, Interval-valued intuitionistic fuzzy soft sets, Decision making, Reduction, Normal parameter reduction.

1 Introduction

Soft set theory was firstly proposed by a Russian Mathematician Molodtsov [1] in 1999. It is a new mathematical tool for dealing with uncertainties, while a wide variety of theories such as probability theory, fuzzy sets [2], and rough sets [3] so on are applicable to modeling vagueness, each of which has its inherent difficulties given in [4]. In contrast to all these theories, soft set theory is free from the above limitations and has no problem of setting the membership function, which makes it very convenient and easy to apply in practice.

Presently, great progresses of study on soft set theory have been made [5, 6, 7]. Furthermore, the soft set model can also be combined with other mathematical models [8, 9, 10]. For example, Soft set models in combination with the interval-valued intuitionistic fuzzy set have been proposed as the concept of the interval-valued intuitionistic fuzzy soft set [11] by jiang et al. At the same time, there are also some efforts which have been done to such issues concerning practical applications [12, 13] of soft sets, especially the employment of soft sets in decision making. Maji et al. [14] first employed soft sets to solve the decision-making problem. Roy et al. [15]

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 642–651, 2011.

presented a novel method of object recognition from an imprecise multi-observer data to deal with decision making based on fuzzy soft sets. Chen et al. [16] pointed out that the conclusion of soft set reduction offered in [14] was incorrect, and then present a new notion of parameterization reduction in soft sets in comparison with the definition to the related concept of attributes reduction in rough set theory. The concept of normal parameter reduction is introduced in [17], which overcome the problem of suboptimal choice and added parameter set of soft sets. However, up to the present, few documents have focused on practical applications of the interval-valued fuzzy intuitionistic soft sets. Actually, a number of real life problems of decision making in engineering, social and medical sciences, economics etc. involve imprecise fuzzy data. The interval-valued fuzzy intuitionistic soft set is a new efficient tool for dealing with diverse types of uncertainties and imprecision embedded in a system. So in this paper, Firstly, we present the algorithm to solve fuzzy decision making problems based on interval-valued intuitionistic fuzzy soft sets, which can help decision makers obtain the optical choice. And then we propose a definition of normal parameter reduction of interval-valued intuitionistic fuzzy soft sets and give a heuristic algorithm, which can delete redundant attributes in decision making process.

This paper is organized as follows. Section 2 reviews the basic notions of interval-valued intuitionistic fuzzy soft sets. Section 3 presents applications of interval-valued intuitionistic fuzzy soft set in a decision making problem, including the algorithm to solve decision making problems based on interval-valued intuitionistic fuzzy soft sets, a definition of normal parameter reduction of interval-valued intuitionistic fuzzy soft sets and a related heuristic algorithm. Section 4 gives an illustrative example. Finally, section 5 presents the conclusion from our work.

2 Basic Notions

In this section, we review some definitions with regard to soft sets, fuzzy soft sets and interval-valued intuitionistic fuzzy soft sets.

Let U be a non-empty initial universe of objects, E be a set of parameters in relation to objects in U, $P(U)$ be the power set of U, and $A \subset E$. The definition of soft set is given as follows.

Definition 2.1 (See [4]). *A pair* (F, A) *is called a soft set over U, where F is a mapping given by*

$$F : A \to P(U) \ . \tag{1}$$

That is, a soft set over U is a parameterized family of subsets of the universe U.

Let U be an initial universe of objects, E be a set of parameters in relation to objects in U, $\xi(U)$ be the set of all fuzzy subsets of U. The definition of fuzzy soft set is given as follows.

Definition 2.2. (See [11]). *A pair* $\left(\tilde{F},E\right)$ *is called a fuzzy soft set over* $\xi\left(U\right)$*, where* $\tilde{F}$ *is a mapping given by*

$$\tilde{F}:E\rightarrow\xi(U)\ . \tag{2}$$

A fuzzy soft set is a parameterized family of fuzzy subsets of *U*, so its universe is the set of all fuzzy sets of *U*.

Atanassov and Gargov [18] first initiated interval-valued intuitionistic fuzzy set (IVIFS), which is characterized by an interval-valued membership degree and an interval-valued non-membership degree.

Definition 2.3 (See [18]). *An interval-valued intuitionistic fuzzy set on a universe X is an object of the form*

$$A=\{\langle x,\mu_A(x),\gamma_A(x)\rangle|x\in X\}\ . \tag{3}$$

where $\mu_A(x): X\rightarrow Int([0,1])$ *and* $\gamma_A(x):X\rightarrow Int([0,1])$ ($Int([0,1])$ *stands for the set of all closed subintervals of [0, 1]) satisfy the following condition:* $\forall x\in X, \sup\mu_A(x)+\sup\gamma_A(x)\leq 1$.

Let *U* be an initial universe of objects, *E* be a set of parameters in relation to objects in *U*, $\zeta(U)$ be the set of all interval-valued intuitionistic fuzzy sets of *U*. The definition of interval-valued intuitionistic fuzzy soft set is given as follows.

Definition 2.4 (See [11]). *A pair* $(\tilde{\varphi},E)$*is called an interval-valued intuitionistic fuzzy soft set over* $\zeta\left(U\right)$*, where* $\tilde{\varphi}$ *is a mapping given by*

$$\tilde{\varphi}:E\rightarrow\zeta\left(U\right)\ . \tag{4}$$

In other words, an interval-valued intuitionistic fuzzy soft set is a parameterized family of interval-valued intuitionistic fuzzy subsets of *U*. Hence, its universe is the set of all interval-valued intuitionistic fuzzy sets of *U*, i.e. $\zeta\left(U\right)$. Since an interval-valued intuitionistic fuzzy soft set is still a mapping from parameters to $\zeta\left(U\right)$, it is a special case of a soft set.

3 Applications of Interval-Valued Intuitionistic Fuzzy Soft Set in a Decision Making Problem

In this section, firstly, we present the algorithm to solve fuzzy decision making problems based on interval-valued intuitionistic fuzzy soft sets, which can help decision makers obtain the optical choice. Secondly, we depict a definition of normal parameter reduction of interval-valued intuitionistic fuzzy soft sets and give a heuristic algorithm, which can delete redundant attributes in decision making process.

3.1 Interval-Valued Intuitionistic Fuzzy Soft Sets Based Algorithm to Solve Fuzzy Decision Making Problems

Actually, a number of real life problems of decision making in engineering, social and medical sciences, economics etc. involve imprecise fuzzy data. The interval-valued fuzzy intuitionistic soft set is a new efficient tool for dealing with diverse types of uncertainties and imprecision embedded in a system. The interval-valued fuzzy intuitionistic soft set is applicable to solve fuzzy decision making problems which involve a large number of imprecise fuzzy data. So we present the algorithm to solve fuzzy decision making problems based on interval-valued intuitionistic fuzzy soft sets, which can help decision makers obtain the optical choice. Before describing this algorithm, we give some related definitions in the following.

Definition 3.1. *For an interval-valued intuitionistic fuzzy soft set* $(\tilde{\varphi},E)$, $U=\{h_1,h_2,\cdots,h_n\}$, $E=\{e_1,e_2,\cdots,e_m\}$, $\mu_{\tilde{\varphi}(e_j)}(h_i)=[\mu^-_{\tilde{\varphi}(e_j)}(h_i),\mu^+_{\tilde{\varphi}(e_j)}(h_i)]$ *is the degree of membership an element* h_i *to* $\tilde{\varphi}(e_j)$.*We denote* $p_{\tilde{\varphi}(e_j)}(h_i)$ *as score of membership degrees for* e_j, *where it is formulated as*

$$p_{\tilde{\varphi}(e_j)}(h_i)=\sum_{k=1}^{n}[(\mu^-_{\tilde{\varphi}(e_j)}(h_i)+\mu^+_{\tilde{\varphi}(e_j)}(h_i))-(\mu^-_{\tilde{\varphi}(e_j)}(h_k)+\mu^+_{\tilde{\varphi}(e_j)}(h_k))]\ . \tag{5}$$

Definition 3.2. *For an interval-valued intuitionistic fuzzy soft set* $(\tilde{\varphi},E)$, $U=\{h_1,h_2,\cdots,h_n\}$, $E=\{e_1,e_2,\cdots,e_m\}$, $\gamma_{\tilde{\varphi}(e_j)}(h_i)=[\gamma^-_{\tilde{\varphi}(e_j)}(h_i),\gamma^+_{\tilde{\varphi}(e_j)}(h_i)]$ *is the degree of non-membership an element* h_i *to* $\tilde{\varphi}(e_j)$. *We denote* $q_{\tilde{\varphi}(e_j)}(h_i)$ *as score of non-membership degrees for* e_j, *where it is formulated as*

$$q_{\tilde{\varphi}(e_j)}(h_i)=-\sum_{k=1}^{n}[(\gamma^-_{\tilde{\varphi}(e_j)}(h_i)+\gamma^+_{\tilde{\varphi}(e_j)}(h_i))-(\gamma^-_{\tilde{\varphi}(e_j)}(h_k)+\gamma^+_{\tilde{\varphi}(e_j)}(h_k))]\ . \tag{6}$$

Definition 3.3. *For an interval-valued intuitionistic fuzzy soft set* $(\tilde{\varphi},E)$, $U=\{h_1,h_2,\cdots,h_n\}$, $E=\{e_1,e_2,\cdots,e_m\}$, $p_{\tilde{\varphi}(e_j)}(h_i)$ *and* $q_{\tilde{\varphi}(e_j)}(h_i)$ *are score of membership and non-membership degree for* e_j, *respectively. We denote* $u_{\tilde{\varphi}(e_j)}(h_i)$ *as score of* h_i *for* e_j, *where it is formulated as*

$$u_{\tilde{\varphi}(e_j)}(h_i) = p_{\tilde{\varphi}(e_j)}(h_i) + q_{\tilde{\varphi}(e_j)}(h_i) \ . \tag{7}$$

Based on these definitions, we give this algorithm shown in Figure 1.

(1) Input an interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi}, E)$ and the parameter set E. $U = \{h_1, h_2, \cdots, h_n\}$, $E = \{e_1, e_2, \cdots, e_m\}$, $\mu_{\tilde{\varphi}(e_j)}(h_i) = [\mu^-_{\tilde{\varphi}(e_j)}(h_i), \mu^+_{\tilde{\varphi}(e_j)}(h_i)]$ is the degree of membership an element h_i to $\tilde{\varphi}(e_j) \cdot \gamma_{\tilde{\varphi}(e_j)}(h_i) = [\gamma^-_{\tilde{\varphi}(e_j)}(h_i), \gamma^+_{\tilde{\varphi}(e_j)}(h_i)]$ is the degree of non-membership an element h_i to $\tilde{\varphi}(e_j)$.

(2) Compute score of membership degrees $p_{\tilde{\varphi}(e_j)}(h_i)$ and score of non-membership degrees $q_{\tilde{\varphi}(e_j)}(h_i)$, for $1 \le i \le n, 1 \le j \le m$, respectively.

(3) Compute the score $u_{\tilde{\varphi}(e_j)}(h_i)$, for $1 \le i \le n, 1 \le j \le m$;

(4) Compute the overall score t_i for h_i such that

$$t_i = u_{\tilde{\varphi}(e_1)}(h_i) + u_{\tilde{\varphi}(e_2)}(h_i) + \ldots + u_{\tilde{\varphi}(e_m)}(h_i)$$

(5) Find k, for which $t_k = \max_{h_i \in U}\{t_i\}$. Then $h_k \in U$ is the optimal choice object.

Fig. 1. Algorithm for decision making on interval-valued intuitionistic fuzzy soft sets

3.2 Normal Parameter Reduction of Interval-Valued Intuitionistic Fuzzy Soft Sets

Obviously, it is not always feasible to only consider the optimal choice about decision in a large number of real applications. Decision makers are more interested in suboptimal choice or ranking of alternatives. Much time is wasted if we make a new decision for data sets in which the data of optimal choice is deleted. Moreover, there are much work to do when new parameters are added to the parameter set. Consequently, it is more reasonable to give a normal parameter reduction of interval-valued intuitionistic fuzzy soft sets which considers the problems of suboptimal choice and added parameters. The normal parameter reduction of interval-valued intuitionistic fuzzy soft sets means rank of alternatives keep invariable after deleting some redundant attributes. It can involve less computation if some data sets are combined or add new parameters and decision makers need suboptimal choice or ranking of alternatives. So the normal parameter reduction is very useful for decision makers. In this section, we propose some definitions and then a heuristic algorithm to achieve the normal parameter reduction of interval-valued intuitionistic fuzzy soft sets, which is based on the above algorithm to solve fuzzy decision making problems.

Definition 3.4. *For an interval-valued intuitionistic fuzzy soft set* $(\tilde{\varphi}, E)$, $U = \{h_1, h_2, \cdots, h_n\}$, $E = \{e_1, e_2, \cdots, e_m\}$, *if there exists a subset* $A = \{e'_1, e'_2, \cdots, e'_g\} \subset E$ *satisfying* $\sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_1) = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_2) = \ldots = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_n)$, *then A is dispensable, otherwise, A is indispensable.* $B \subset E$ *is defined as a normal parameter reduction of E, if the two conditions as follows are satisfied*

(1) B is indispensable

(2) $\sum_{e_k \in E-B} u_{\tilde{\varphi}(e_k)}(h_1) = \sum_{e_k \in E-B} u_{\tilde{\varphi}(e_k)}(h_2) = \ldots = \sum_{e_k \in E-B} u_{\tilde{\varphi}(e_k)}(h_n)$

Based on the above definition, we give the normal parameter reduction algorithm as follows:

```
(1)  Input an interval-valued intuitionistic fuzzy soft set
     $(\tilde{\varphi}, E)$ and the parameter set E.
(2)  Compute score of membership degrees $p_{\tilde{\varphi}(e_j)}(h_i)$ and score of
     non-membership degrees $q_{\tilde{\varphi}(e_j)}(h_i)$, for $1 \le i \le n, 1 \le j \le m$,
     respectively.
(3)  Compute the score $u_{\tilde{\varphi}(e_j)}(h_i)$, for $1 \le i \le n, 1 \le j \le m$;
(4)  Check A, where $A = \{e'_1, e'_2, \cdots, e'_g\} \subset E$, if
     $\sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_1) = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_2) = \ldots = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_n)$,
     and then A is put into a candidate parameter reduction set.
(5)  Find the maximum cardinality of A in the candidate
     parameter reduction set.
(6)  Get E-A as the optimal normal parameter reduction.
```

Fig. 2. Algorithm for normal parameter reduction of interval-valued intuitionistic fuzzy soft sets

4 Example

In this section, in order to explicitly clarify the above algorithms, the following example is given.

Example 4.1. Let $(\tilde{\varphi}, E)$ be an interval-valued intuitionistic fuzzy soft set with the tabular representation displayed in Table 1. Suppose that

$$U = \{h_1, h_2, h_3, h_4, h_5, h_6\} \text{ and } E = \{e_1, e_2, e_3, e_4, e_5\}.$$

Table 1. an Interval-Valued Intuitionistic Fuzzy Soft set $(\tilde{\varphi}, E)$

U/E	e_1	e_2	e_3	e_4	e_5
h_1	[0.6,0.7],[0.0,0.2]	[0.3,0.4],[0.4,0.5]	[0.5,0.7],[0.0,0.2]	[0.4,0.5],[0.4,0.5]	[0.4,0.6],[0.2,0.3]
h_2	[0.2,0.4],[0.3,0.5]	[0.5,0.6],[0.2,0.3]	[0.3,0.5],[0.1,0.3]	[0.7,0.8],[0.0,0.1]	[0.2,0.4],[0.4,0.5]
h_3	[0.3,0.6],[0.1,0.3]	[0.6,0.7],[0.2,0.3]	[0.7,0.9],[0.0,0.1]	[0.5,0.7],[0.1,0.3]	[0.1,0.3],[0.3,0.6]
h_4	[0.4,0.5],[0.1,0.3]	[0.6,0.8],[0.0,0.1]	[0.2,0.5],[0.1,0.3]	[0.3,0.5],[0.2,0.4]	[0.1,0.2],[0.6,0.7]
h_5	[0.5,0.8],[0.0,0.2]	[0.7,0.8],[0.0,0.2]	[0.7,0.9],[0.0,0.1]	[0.6,0.7],[0.1,0.2]	[0.0,0.2],[0.5,0.7]
h_6	[0.6,0.8],[0.0,0.2]	[0.4,0.6],[0.1,0.2]	[0.4,0.7],[0.0,0.2]	[0.5,0.6],[0.3,0.4]	[0.3,0.4],[0.5,0.6]

According to the algorithm for decision making on interval-valued intuitionistic fuzzy soft sets, we can compute score of membership degrees $p_{\tilde{\varphi}(e_j)}(h_i)$ for $(\tilde{\varphi}, E)$, which is shown in Table 2.

Table 2. The score of membership degrees for $(\tilde{\varphi}, E)$

U/E	e_1	e_2	e_3	e_4	e_5
h_1	1.4	-2.8	0.2	-1.4	2.8
h_2	-2.8	-0.4	-2.2	2.2	0.4
h_3	-1	0.8	2.6	0.4	-0.8
h_4	-1	1.4	-2.8	-2	-1.4
h_5	1.4	2	2.6	1	-2
h_6	2	-1	-0.4	-0.2	1

We can compute score of non-membership degrees $q_{\tilde{\varphi}(e_j)}(h_i)$ for $(\tilde{\varphi}, E)$, which is shown in Table 3.

According to interval-valued intuitionistic fuzzy soft sets based algorithm to solve fuzzydecision making problems, we have $t_1 = -1$, $t_2 = -4$, $t_3 = 3.2$, $t_4 = -7.6$, $t_5 = 8$ and $t_6 = 1.4$. Hence h_5 is the best choice because $t_5 = \max_{h_i \in U}\{t_i\}$.

Table 3. The score of non-membership degrees for $(\tilde{\varphi}, E)$

U/E	e_1	e_2	e_3	e_4	e_5
h_1	1	-2.9	0.2	-2.4	2.9
h_2	-2.6	-0.5	-1	2.4	0.5
h_3	-0.2	-0.5	0.8	0.6	0.5
h_4	-0.2	1.9	-1	-0.6	-1.9
h_5	1	1.3	0.8	1.2	-1.3
h_6	1	0.7	0.2	-1.2	-0.7

We can compute the score for $(\tilde{\varphi}, E)$, which is shown in Table 4.

Table 4. The score for $(\tilde{\varphi}, E)$

U/E	e_1	e_2	e_3	e_4	e_5	t_i
h_1	2.4	-5.7	0.4	-3.8	5.7	-1
h_2	-5.4	-0.9	-3.2	4.6	0.9	-4
h_3	-1.2	0.3	3.4	1	-0.3	3.2
h_4	-1.2	3.3	-3.8	-2.6	-3.3	-7.6
h_5	2.4	3.3	3.4	2.2	-3.3	8
h_6	3	-0.3	-0.2	-1.4	0.3	1.4

According to the proposed algorithm for normal parameter reduction of interval-valued intuitionistic fuzzy soft sets, from Table 5, we can obtain $\{e_2, e_5\}$ satisfying

$$\sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_1) = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_2) = \ldots = \sum_{e_k \in A} u_{\tilde{\varphi}(e_k)}(h_6) = 0.$$

Thus $\{e_1,e_3,e_4\}$are the normal parameter reduction of the interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi},E)$. It means the rank of alternatives can not be changed after deleting attributes of e_2 and e_5.

5 Conclusion

Some work on the interval-valued intuitionistic fuzzy soft sets has been done by Jiang et al. They introduced the concept of the interval-valued intuitionistic fuzzy soft sets and the complement, "and", "or", union, intersection, necessity and possibility operations. Furthermore, the basic properties of the interval-valued intuitionistic fuzzy soft sets are also presented and discussed. However, up to the present, few documents have focused on such issues concerning practical applications of interval-valued intuitionistic fuzzy soft sets. So in this paper, we present the algorithm to solve decision making problems based on interval-valued intuitionistic fuzzy soft sets which is one of practical applications. And then we propose a definition of normal parameter reduction of interval-valued intuitionistic fuzzy soft sets and give a heuristic algorithm to achieve the normal parameter reduction of interval-valued intuitionistic fuzzy soft sets, which considers the problems of sub-optimal choice and added parameters. Finally, an illustrative example is employed to show the validity of our algorithms on decision making and normal parameter reduction of interval-valued intuitionistic fuzzy soft sets.

References

1. Molodtsov, D.: Soft set theory_First results. Computers and Mathematics with Applications 37(4/5), 19–31 (1999)
2. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
3. Pawlak, Z.: Rough sets. International Journal Information Computer Science 11, 341–356 (1982)
4. Molodtsov, D.: The Theory of Soft Sets. URSS Publishers, Moscow (2004) (in Russian)
5. Ali, M.I., Feng, F., Liu, X., Min, W.K., Shabira, M.: On some new operations in soft set theory, Comput. Comput. Math. Appl. 57(9), 1547–1553 (2009)
6. Pei, D., Miao, D.: From soft sets to information systems. In: The proceeding of 2005 IEEE International Conference on Granular Computing. IEEE GrC 2005, pp. 617–621. IEEE Press, Los Alamitos (2005)
7. Herawan, T., Deris, M.: A direct proof of every rough set is a soft set. In: Proceeding of the Third Asia International Conference on Modeling and Simulation, pp. 119–124. AMS, Bali (2009)
8. Feng, F., Li, C., Davvaz, B., Ali, M.I.: Soft sets combined with fuzzy sets and rough sets: A tentative approach. In: Soft Computing - A Fusion of Foundations. Methodologies and Applications, pp. 899–911. Springer, Heidelberg (2009)
9. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9(3), 589–602 (2001)
10. Maji, P.K., Biswas, R., Roy, A.R.: Intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 9(3), 677–692 (2001)

11. Jiang, Y., Tang, Y., Chen, Q., Liu, H., Tang, J.: Interval-valued intuitionistic fuzzy soft sets and their properties. Computers and Mathematics with Applications 60(3), 906–918 (2010)
12. Zou, Y., Xiao, Z.: Data analysis approaches of soft sets under incomplete information. Knowl.-Based Syst. 21(8), 941–945 (2008)
13. Herawan, T., Mat Deris, M.: A Soft Set Approach for Association Rules Mining. Knowledge Based Systems (2010) doi: 10.1016/j.knosys.2010.08.005
14. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
15. Roy, A.R., Maji, P.K.: A fuzzy soft set theoretic approach to decision making problems. Computers and Mathematics with Applications 44, 1077–1083 (2002)
16. Chen, D., Tsang, E.C.C., Yeung, D.S., Wang, X.: The parameterization reduction of soft sets and its applications. Computers and Mathematics with Applications 49(5-6), 757–763 (2005)
17. Kong, Z., Gao, L., Wang, L., Li, S.: The normal parameter reduction of soft sets and its algorithm. Computers and Mathematics with Applications 56(12), 3029–3037 (2008)
18. Atanassov, K., Gargov, G.: Interval valued intuitionistic fuzzy sets. Fuzzy Sets and Systems 31(3), 343–349 (1989)

Elasticity Study: DBrain Project Grid-Cloud Integration

Dani Adhipta, Mohd Fadzil Hassan, and Ahmad Kamil Mahmood

Computer and Information Sciences Department,
Universiti Teknologi PETRONAS, Perak, Malaysia
dani@ugm.ac.id, mfadzil_hassan@petronas.com.my,
kamilmh@kperak.com.my

Abstract. DBrain is a collaborative project involving multi-discipline fields aimed primarily to provide information and assistance to patient-physician interaction on Dementia disease. This paper explores the possibility and feasibility of the Grid implementation over the Cloud infrastructure to extend and enhance the current resources capacity for catering future expansion. The paper presented a study on the Cloud's elasticity to achieve effective operational cost within the context of DBrain project. Finer details of the current metric is explored and extended for the elasticity indicator.

Keywords: DBrain Project, Grid and Cloud Integration, Elasticity.

1 Introduction

The DBrain project is an ongoing collaborative project aimed to provide information of Dementia disease which includes assistance to the patients and physicians, family members who carries inherited gene, even to the caregiver to track or monitor family members who are suffering from Dementia in real-time. This project was initiated in the early 2009 involving few academic and private institutions in Malaysia providing different specialized services, as highlighted in Figure 1.

The infrastructure is designed and built as a complete health information system solution ranging from the web services portal, high performance computing (HPC) and database records, patient status tracking and monitoring services. The HPC for massive computation is the focus of this paper to efficiently fulfill an increase in users' demands and needs, hence it only makes sense to explore the possibility and feasibility of grid-cloud integration and most importantly utilize the cloud's elasticity capability in order to achieve the operational cost-effective for the project.

This cloud's elasticity study is very important for the future's consideration in supplementing the current infrastructure which has been already established and running. The cloud's technical properties, such as conformity and compatibility with the grid's middleware, are the also the focus of this paper. The current implementation, on-demand elasticity for the dynamic scalability, *i.e.* ability to grow and shrink according to the need, is considered of utmost importance to guarantee a degree of service level according to the agreement (SLA).

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 652–662, 2011.

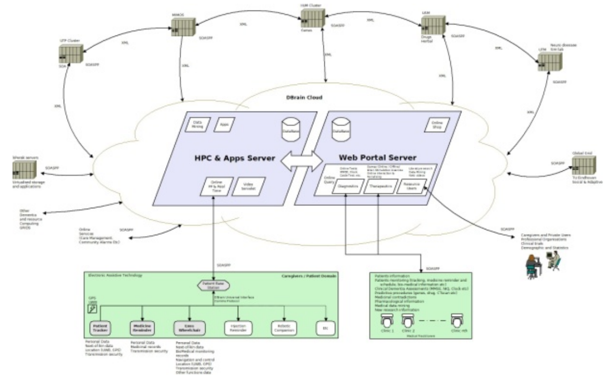


Fig. 1. DBrain Project Architecture

The DBrain current implementation employs two major Grid middleware, Globus Toolkit and gLite for heterogeneous environment, *i.e.* Linux operating system under different versions and hardware (32-bit and 64-bit), in order provide massive computation involving protein folding, pseudo-gene sequencing, and molecular docking along with very large database records searches as shown in Figure 2. Since this project is still at the prototyping stage, only few users are eligible to access and submit concurrent jobs.

The current infrastructure within the DBrain virtual organization (VO) consists of 6 different sites utilizing 24 primary servers with a total of 120 CPUs (or cores).

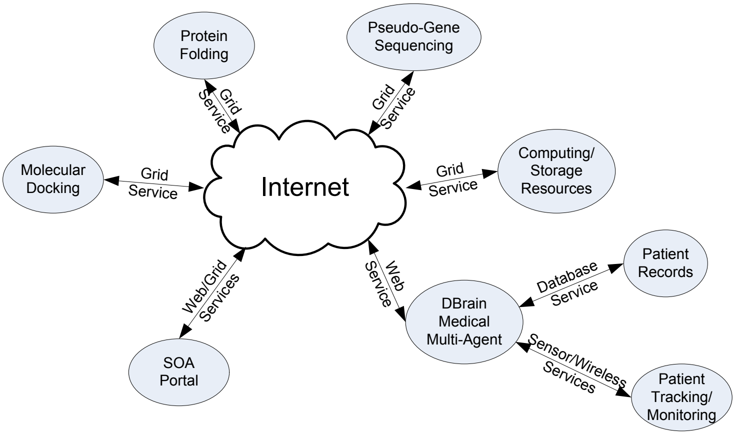


Fig. 2. DBrain collaborative environment

This paper will explore and evaluate the on-demand service provisioning which exhibits dynamic scalability and flexibility, *i.e.* the elasticity. It will also discuss on how the objective to achieve cost-effective estimation could be met.

2 Grid-Cloud Integration

At present, there is no well-established standard for the cloud scientific computing; it is still an ongoing progress as more communities are drawn into this field. However, performance wise , the latency for large data transfer is still an issue and the project reported in [1] tried to improve the technology. In other words, an established guideline must be provided as reported in [2], perhaps even the testing procedure for the grid-cloud integration is required as well.

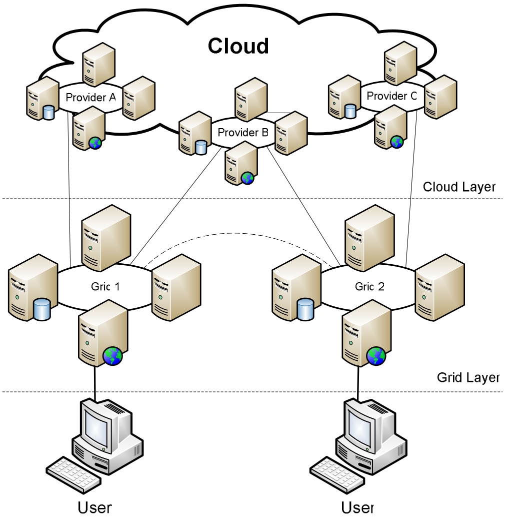


Fig. 3. Grid-Cloud Integration

By definitions, Grid and Cloud span across multiple different geographically distributed sites in heterogeneous environment. However in Grid, there is no central control or management while Cloud tends to be the opposite since the infrastructure provided by a single company will most likely have homogenous resources. Furthermore, by employing multiple different Clouds' providers, this inevitably may also fall into decentralized and heterogeneous category inevitably as well. The nature of Grid computing as a high performance distributed computing is to share its resources for remote job execution to solve large scale and massive problems. In the DBrain project, the Grid primarily provide services for the computation of protein folding, pseudo-gene sequencing, and molecular docking, which all tasks fall under the category of biotechnology and biochemistry problem solving.

2.1 Grid Middleware

The most common Grid middleware being implemented are the Globus Toolkit and gLite, and DBrain VO employs both since each institution is independent of its individual implementation preferences. The Grid HPC as a scientific computing is computing intensive by native and tends to be sensitive when it comes to delay; hence, this latency in communication has yet to be studied.

The basic Grid middleware provides functionalities such as resource management, monitoring and discovery, and security framework for establishing communication with the users or other Grid infrastructures. The current standard already established for the Grid is the service oriented architecture (SOA) open grid services architecture (OGSA) which covers web services resources framework (WSRF) using web services definition language (WSDL). OGSA is a further improvement and extension over open grid services infrastructure (OGSI) to standardized the interoperability property [3].

The current traditional Grid works in batch and queue processing system rather than instantaneous real-time interaction for scientific computing. Commonly the Grid middleware requires a static set of well-defined resources and relies on a reservation scheme for users to utilize.

The future Grid architecture must conform to the representational state transfer (REST) style [4] by adopting the stateless communication to provide even higher site independent, fault tolerance capacity and at the end, better dynamic scalability. In other words, pushing the ability to decouple sites interdependencies for the tightly coupled infrastructure is the goal to achieve higher performance.

2.2 Cloud Technology

In relation to Grid, Cloud computing technology offers technical characteristics of loosely coupled interaction, hence it is expected that in the occurrence of hardware or site failure, only reduced functionalities will be experienced. In other words, resources provision dysfunctionality will normally be avoided.

Cloud computing tends to have user-centric functionalities and services for building customized computing environments. It is more inclined towards industry oriented and follows the application-driven model.

Inherently, when a third party is entrusted for running the Cloud as an infrastructure service, there are always some concerns. For example given, data storage, and the following factors must be well understood and covered by service license agreement (SLA):

- bandwidth performance
- sharing physical storage
- availability
- privacy and confidentiality

Since the Grid natively requires a static set of well-defined resources and relies on a reservation scheme as mentioned previously, the Cloud's infrastructure as a service (IaaS) could be considered the most appropriate solution. IaaS has the highest flexibility and offers processing, storage, and network capacity as the complete provision set. Hence the provider may be considered as hardware provider, in other words this would be called hardware as a service (HaaS).

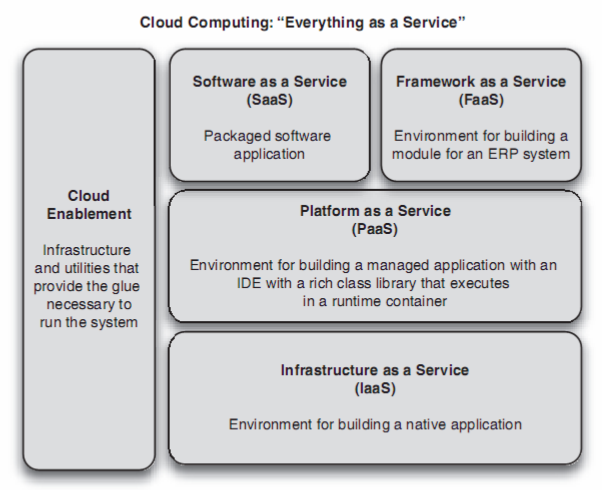

Cloud Computing: "Everything as a Service"
Cloud Enablement
Infrastructure and utilities that provide the glue necessary to run the system
Software as a Service (SaaS)
Packaged software application
Framework as a Service (FaaS)
Environment for building a module for an ERP system
Platform as a Service (PaaS)
Environment for building a managed application with an IDE with a rich class library that executes in a runtime container
Infrastructure as a Service (IaaS)
Environment for building a native application

Fig. 4. Cloud Services provision [5]

2.2.1 Virtualization vs. Emulation

Virtual System Environments (VSE) within an administrative domain and management for the typical High Performance Computing (HPC) environment for performing the biochemistry massive computation is constrained by the operating system (OS) together with the run-time environment (RTE). Such example when running a parallel application, MPI, will require cloud virtualization platform rather than emulation platform.

Utilization of virtualization layer offers flexibility of Grid implementation over the Cloud as compared to emulation in which only a single application could be run at a time in the native environment. Virtualization layer extends the physical infrastructure through hardware independence which can be in the form of isolation or encapsulation to provide flexibility for the heterogeneous software stacks. At the end, capacities such as elasticity (or resources scaling), workload balance, dynamic partitioning and resizing are available when virtualized [6].

Goldberg's definition of virtualization is to allow guest operating system running at the hardware level while emulation is only allowing part of the application code, *i.e.* the microcode, running via interface (*e.g.* API) and not as the physical resources [7]. Furthermore, by this definition for the Grid-cloud integration, virtualization is the more appropriate solution rather than emulation to provide the desired functions.

2.2.2 Cloud Middleware Comparison

Given the need to study and evaluate of the Grid-cloud integration possibility and feasibility, three cloud middleware that support Grid have been investigated. It is the aim to employ off-the-shelf cloud opensource-based middleware, just as Globus Toolkit and gLite middleware which are currently adopted for the Grid environment. Therefore three candidates namely Eucalyptus, OpenNebula, and Nimbus cloud middleware have been evaluated. A previous comparison has been reported in [8].

However, the DBrain project requirement dictates that the cloud middleware should support at least one of the Grid middleware by design; hence Nimbus and OpenNebula are worth evaluating further since these Grid-Cloud technologies are the most promising for the future infrastructure expansion. Their primary strength lays on their deployment simplicities, virtual platforms performance, compatibilities with Globus and gLite grid middleware, and the X509 credential-based security, file transfer support via GridFTP or even S3 Cumulus. Nimbus stands out better since it is designed more toward cooperative scientific community.

As a point of reference, Eucalyptus is compatible with Amazon's EC2 interface and is designed more towards web-based applications for corporate enterprise computing. However, it also supports, although not specifically, IaaS for the Grid via additional client-side interfaces.

In general, most cloud infrastructure is a dynamic infrastructure which consists of elastic web servers, elastic application servers, and elastic database servers. It has the following main characteristics [9]:

- Automatic scalability and load-balancing, failover in terms of virtualization.
- Supports global resource sharing through the internet.

2.2.3 OpenNebula

OpenNebula manages and provides the separation of resources from the service management. By this design, [6] describes OpenNebula having the benefits in the following:

- The dynamic capacity of the cluster, to be deployed then activated or shutdown after deactivation is on demand for both local physical and remote virtual resources.
- The physical resources of the provider could be used to execute worker nodes bound to different cluster-based services, thus isolating their workloads and partitioning the performance assigned to each one.
- The virtual worker nodes of a service can have multiple heterogeneous software configurations. The service images follow "install once and deploy many" approach, which will reduce operational cost.

2.2.4 Nimbus

By design, Nimbus exposes WSRF interface which enable Globus Toolkit grid middleware support and elastic cloud computing (EC2) interface for the self-configuring virtual cluster. Certainly it provides the IaaS for the Grid-Cloud integration as well. The shortcoming of Nimbus is that there is no provision for dynamic allocation and load balancing of computing resources for the virtual nodes and the dynamic scaling of virtual clusters using resources from remote cloud providers.

2.3 Grid vs. Cloud

In summary for the Grid-Cloud integration, it is apparent that these distributed technologies are aimed for different purposes although they are inherently exhibit almost identical characteristics. Grid is utilized specifically for science computing while Cloud is more toward business/commercial corporate infrastructure. Therefore, in the future most likely there will be well-established standards to interface the two technologies for cost saving science computing.

Table 1. Eucalyptus vs. OpenNebula vs. Nimbus Cloud [8]

	Eucalyptus	OpenNebula	Nimbus
philosophy	mimic Amazon EC2	private, highly customizable	resources tailored to scientific need
customizability	partially for administrator, minimum for user	almost everything	many parts except for image storage and Globus credentials
DHCP	only on cluster controller	variable	on individual compute node
internal security	tight, still requires many root access	looser, can be tightened if required	fairly tight, unless fully private cloud is intended
user security	web based interface credential	login to head node	utilizes registered user X509 credential
network issue	DHCP on cluster controller	manually configured with many options	DHCP on every node with assigned MAC

Table 2. Cloud vs. Grid [10]

Characteristics	Cloud	Grid
service oriented	yes	yes/no
loose coupling	yes	yes/no
fault tolerant	yes	yes/no
business model	yes	yes/no
ease of use	yes	yes/no
TCP/IP	yes	yes/no
high security	yes	yes
virtualization	yes	yes

3 Cloud Elasticity

In-depth analysis of the cloud as elastic site has been published by [11]. The investigation focuses on Nimbus-based cloud between three institutions namely University of Chicago, Indiana University, and Amazon EC2. However for the on-demand policy, only very basic logic is utilized, simply by adding a virtual node when there is a new job queued, and terminates the node when the queue is empty. The most important results that may contribute to the DBrain project is the optimum duration required for the deployment to take place, from initialization until the time when services and application could be executed. These results will serve as a very good estimation in the future for DBrain Grid-Cloud integration.

Any action taken at the cloud site, be it the elasticity or the management, is transparent if its sub-action is automatic and implicitly performed [12]. This transparency includes the process of configuration and reconfiguration according to the application as well.

3.1 Elasticity and Performance

The MPEG decoding case, for example at Amazon EC2 cloud, has been studied and the result of the study highlights the gap that exists between the expected and actual performances. This is due to the single-threaded libraries which are unable to take advantage of CPU's multi-core capability while the network latency prevents the linearity of speedup. The performance or level of service is define by the following equations used to estimate the number of nodes in CPU (or core) unit in terms of workload and time variables [13].

$$cores = \sum_{i=0}^{n} c(i) \tag{1}$$

$$load = \frac{\sum_{i=0}^{n}\sum_{j=0}^{c(i)} l(j)}{cores} = \frac{\sum_{i=0}^{n}\sum_{j=0}^{c(i)} l(j)}{\sum_{i=0}^{n} c(i)} \tag{2}$$

with
n=number of nodes
$c(i)$=number of CPU cores for i node
$l(j)$=load for CPU core

3.2 Metric Indicator

We explore and propose a simple indicator to measure the elasticity of DBrain Grid-Cloud integration given the time when the computing resources must be scaled up or down transparently (automatically) according the running process demand. The primary parameters which may be significantly accounted are the load level (in percent of cumulative CPU loads), queue (the jobs waiting for processing), and the nodes number for scaling up or down. Then assumptions can be created when the load is no longer able to be handled at current time by giving a maximum level of 95% as the full load.

For example, when there is maximum load and the queue grows longer, then it becomes an indicator to scale up the computing resources (by adding more virtual nodes). In this case study the queue length is set to 10 and the nodes expansion is set to 2 nodes, as an illustration for the DBrain cloud elasticity. This scaling up logic is defined in the following:

```
if $load \geq 95$
and $queue \geq 10$
then $nodes = nodes_{current} + 2$
```

The opposite applies to the similar logic when the load and queue has decreased by reducing 1 virtual node until finally reaching the minimum given nodes number. In other words, no further scaling down is performed when the minimum number of virtual node is reached.

```
if load<95
and queue<10
then nodes=nodes_current-1
```

However, further evaluation is required to decide whether to expand or shrink the virtual nodes when it occurs at the threshold. This is defined below:

```
if load=95
and queue<10
then wait and evaluate
```

The 3-stages of scaling indicator are in Table 3.

Table 3. Metric Elasticity Indicator

Load (%)	Queue	Elasticity (nodes #)	trigger
<95	<10	0/1	No/scale-down
95	<10	X	Evaluate
≥95	≥10	2	Yes/scale-up

When scaling down is needed, the elasticity factor is only set to 1 node. This is to anticipate if there will be another increase in load and queue again in the very near future. Otherwise stated, a graceful degradation is highly desired characteristic.

This simple straight forward elasticity indicator may be sufficient as a trigger to add or reduce the number of cloud virtual nodes. However for finer grain logic, more complex job scheduling technique can be employed as well which may consist of queue duration, job priority, etc [14]. For example when there is a job that has high priority enters the queue and the load has already reached the maximum, without calculating the queue length then scaling up can be performed.

4 Conclusions

This paper has presented the ongoing DBrain project with a discussion on the possible implementation and future expansion aspects. The cloud as the virtual infrastructure offers the possible and feasible strategy in catering cost-efficient services to the DBrain project community. Previous research works and implementation experiences have proved to be invaluable information to produce what-if scenarios, particularly when elasticity is concerned in dealing with dynamic scalability.

The user perceived access response is the very important aspect of the service provision. The infrastructure elasticity at an acceptable performance and operational cost-effective will impact the sustainability of the whole initiative in the long-run. On top of this, successful internet-based business model depends on the efficient implementation in terms of operating cost and the technical services; hence user demands which sometimes lead to over-provisioning during the peak periods must be avoided.

In the future, extending the current Grid resources broker and job scheduler, metascheduler such as Nimrod-G and Condor-G [15] may be studied and implemented to achieve finer grain Grid-Cloud workload distribution. Hence as a conclusion, the DBrain project SOA-based for the Grid-Cloud integration may evolve to become what is so called Enterprise Service Oriented Architecture (ESOA) [9] in a hybrid environment when possibly dictated by demands.

Acknowledgments. This paper was submitted under the Techno Fund of DBrain Project Malaysia grant and support from the Universiti Teknologi PETRONAS, 2009-2011.

References

1. Wang, L., Tao, J., Kunze, M., Castellanos, A.C., Kramer, D., Karl, W.: Scientific Cloud Computing: Early Definition and Experience. In: 10th IEEE International Conference on High Performance Computing and Communications, pp. 825-830 (2008)
2. Rings, T., Grabowski, J., Schulz, S.: On the Standardization of a Testing Framework for Application Deployment on Grid and Cloud Infrastructures. In: 2nd International Conference on Advances in System Testing and Validation Lifecycle, IEEE Computer Society Press, Los Alamitos (2010)
3. McMullen, D.F., Devadithya, T., Chiung, K.: Integrating Instruments and Sensors into the Grid with CIMA Web Services. In: 3rd APAC Conference on Advanced Computing, Grid Applications and e-Research, APAC 2005 (2005)
4. Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. ACM Transactions on Internet Technology 2(2), 115–150 (2002)
5. Rosenberg, J., Mateos, A.: The Cloud at Your Service. In: The Cloud at Your Service, p. 247. Manning Publications, New York (2011)
6. Moreno-Vozmediano, R., Montero, R.S., Llorente, I.M.: Elastic Management of Cluster-based Services in the Cloud. In: ACDC 2009, Barcelona, Spain (2009)
7. Gallard, J., Lèbre, A., Vallée, G., Morin, C., Gallard, P., Scott, S.L.: Refinement proposal of the goldberg's theory. In: Hua, A., Chang, S.-L. (eds.) ICA3PP 2009. LNCS, vol. 5574, pp. 853–865. Springer, Heidelberg (2009)
8. Sempolinski, P., Thain, D.: A Comparison and Critique of Eucalyptus, OpenNebula and Nimbus. In: IEEE International Conference on Cloud Computing Technology and Science. University of Notre Dame (2010)
9. Tang, L., Dong, J., Zhao, Y., Zhang, L.J.: Enterprise Cloud Service Architecture. In: 3rd IEEE International Conference on Cloud Computing (2010)
10. Gong, C., Liu, J., Zhang, Q., Chen, H., Gong, Z.: The Characteristics of Cloud Computing. In: 39th International Conference on Parallel Processing Workshops. Department of Computer Sciences, National University of Defense Technology, Changsa, China (2010)
11. Gong, C., et al.: The Characteristics of Cloud Computing. In: 39th International Conference on Parallel Processing Workshops, Department of Computer Sciences, National University of Defense Technology, Changsa, China (2010)
12. Marshall, P., Keahey, K., Freeman, T.: Elastic Site: Using Clouds to Elastically Extend Site Resources. In: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, Melbourne, Australia (2010)

13. Gallard, J., Vallee, G., Naughton, T., Lebre, A., Scott, S.L., Morin, C.: Architecture for the Next Generation System Management Tools for Distributed Computing Platforms. INRIA, 7325, inria-00494328 (2010)
14. Dornemann, T., Juhnke, E., Freisleben, B.: On-Demand Resource Provisioning for BPEL Workflows Using Amazon's Elastic Compute Cloud. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. IEEE, CCGRID (2009)
15. Peixoto, M.L.M., Santana, M.J., Estrella, J.C., Tavares, T.C., Kuehne, B.T., Santana, R.H.C.: A Metascheduler Architecture to provide QoS on the Cloud Computing. In: 17th IEEE International Conference on Telecommunications (ICT), p. 650. IEEE, Los Alamitos (2010)
16. Vivekanandan, K., Ramyachitra, D.: A Study on Scheduling in Grid Environment. International Journal on Computer Science and Engineering (IJSE) 3(2), 940–950 (2011)

Embedded Backups for Link Failure Recovery in Shortest Path Trees

Muhammad Aasim Qureshi, and Mohd Fadzil Hassan

Computer and Information Sciences Department
Universiti Technologi PETRONAS
Perak, Malaysia
maasimq@hotmail.com, mfadzil_hassan@petronas.com.my

Abstract. Shortest Path Tree Problem has always been a popular problem but with the devise of Dijkstra's algorithm, SPT and many related problems received immense intention of researchers. Shortest Path Tree plays an important role in many applications like robot navigation, games, transportation and communication routing, etc. In many applications like network and vehicle routing, fast and reliable recovery from the failure is desired. Recovery from these failed links need means and/or plan, with least additional healing cost for the prolongation of the process with no or minimum delay. This paper presents an approach to recover from undesirable state of link failure(s) back to the normal state of working. A shortest path tree is being extracted from the graph with alternate path at each point (junction) keeping the cost as low as possible.

Keywords: shortest path algorithm, recovery, alternate shortest path, multiple links failure recovery.

1 Introduction

Robot navigation, communication and vehicle routing are important areas in computer science which, need robust and reliable processing. In these areas of research, the most optimal path that lead to safe and reliable accomplishment of job (i.e. packet transfer in communication, vehicle and robot movement to destinations in the later two) is desirable. The theoretical aspect of the solutions for these and many other practical problems fall under the area of shortest path in theoretical computer science.

In graph theory, the shortest path problem is the problem of finding an optimized path between two vertices (or nodes) such that the sum of the weights of its constituent edges is minimized. For an example, finding the quickest way from one location to another on a road map; in this case, the vertices represent locations and the edges represent segments of road and are weighted by the time needed to travel that road segment.

Formally, given a weighted graph G(V, E),where V is a set of all vertices and E is set of all edges. Shortest path tree problem is to find a tree of paths rooted at source node with shortest path to the other vertices from the source.

Due to its innate nature and wide range of applications, shortest path algorithms plays an important role in solving many optimization problems like robot navigation,

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 663–671, 2011.

games, vehicle routing and networks. In fact, the shortest path problem is extensively applied in communication, computer systems, transportation networks and many other practical problems [1]. Shortest path algorithms also support the claim of the shortest processes for development [2][3]. Its solution leads to the efficient management of control workflow processes [4][5]. In short many problems can be mapped, directly or indirectly to the shortest path problem.

Single source shortest path is one of the challenging areas of theoretical computer science [6][7] which is based on the Dijkstra's algorithm [8], presented in 1959. There are many variations of shortest path problem; Hershberger discussed the difficulty of different algorithms in [9]. In short many authors like Seth Pettie [10], Thorup [11][12], Raman [13] and many others have contributed in one way or the other to the solution(s) of this problem.

Link failure is a common problem in communication and vehicle routing problems which can lead to system crash. In order to recover from such failures, systems are made capable to adopt a new path and continue execution. Recovery from such link failure(s) at run time is one of the enormous Single Source Shortest Path (SSSP) based problems. Its applications can be seen in transportation networks as well as in communication networks. This problem can formally be defined as follows:

Given a graph G(V,E), where V is a set of vertices and E is a set of edges. $\boldsymbol{G}$ is an undirected weighted 2-edge-connected graph having no negative weight edge, with pre-specified source vertex '$\boldsymbol{s}$' and destination vertex '$\boldsymbol{t}$' such that $\boldsymbol{s},\boldsymbol{t} \in$ V. Let $\mathrm{Cost}\ (\mathrm{PG}(\boldsymbol{s},\boldsymbol{t}))$ be the cost of the Path $\boldsymbol{P}$ in graph $\boldsymbol{G}$ from s to $\boldsymbol{t}$. We have to find a tree T = { t_1, t_2, t_3, ... , t_k }, where $t_i \in$ E for all i = 1 to k, rooted at s, that have alternate shortest paths on the failure of any link on shortest path tree(from the very point of failure).

A $\mathrm{O(n)}$ times algorithm was presented in [14] specifically limited to planar undirected graphs while the same paper presents a $\mathrm{O(m + n.\alpha(2n, n))}$ times algorithm for undirected graphs with integral edge weights. A. M. Bhosle suggested, in [14][15], that the solution to a single link recovery problem can be used for alternate path routing mechanism that is used in ATM networks. Yueping Li, in [16], presented an $\mathrm{O(mn + mn^2 \log n)}$ time optimal shortest path tree that plays an important role in a single link failure recovery.

Complete details of the algorithm along with the explanations and discussions are provided in sections 2. Sections 3 include a working example showing the usability with a step by step execution of the algorithm. Algorithm is analyzed in full detail in section 4 where its complexity is proven. Finally, in section 5, concluding remarks are given.

2 Algorithm

This paper presents a backup schema for an inevitable collapse(s) during a traversal of a shortest path. The focus of the work is to provide a recovery from multiple link failures in a row with an attempt on a single failure at a time. It presents the best alternate path with the most optimized one so that, in case of a direct link between two nodes collapses or fails, the process can continue through alternate path. This effort is to provide a mechanism for the continuation of the progression in the face of failure through backup path(s). With this provision of back up path (alternate paths), shortest path can be traversed successfully from source to destination with the capability of facing multiple failures on the way.

This work is a continuation of the works presented in [17][18]. In order to provide backup flexibility, all scenarios need to be covered. Alternate paths from every vertex that lie on the shortest path tree needs to be calculated and entrenched into the Shortest Path Tree (SPT) including the worst one. It is needed to keep track of the shortest path from all the vertices to destination vertex. Once these shortest paths have been calculated, the two edges from every vertex will be computed—best and second best, connecting previously calculated shortest paths. The second best path will be providing backup for the best path.

The algorithm is logically divided into two sequential steps. Step I calculates the shortest path from all vertices to destination and step II on the basis of the results generated in step I, calculates shortest path by running a Breadth First Search (BFS) like algorithm to calculate a set of best minimal paths as well as alternate path. Full details related to each step are as below:

2.1 Step I—Calculating Shortest Path from All Vertices to Destination

In any undirected graph

$$\mathbf{P_{a,b} = P_{b,a}}$$

Where $\mathbf{P_{x,y}}$ stands for path—direct or indirect, between x and y.

$$\mathbf{T_{all,t} = SPT_{t,all}}$$

Where $T_{all,t}$ is tree of all shortest paths to destination while $\mathbf{SPT_{t,all}}$ is shortest path tree from source to destination.

So we need a shortest path algorithm to calculate such tree. There are algorithms available with their own strengths and weaknesses. This research will be adopting the Two-Phase SPT algorithm presented in [19][20] due its efficiency and reliability (its worst case complexity is the same as that of Dijkstra's but on average and best case it is better. In addition, it is easy to understand and implement with simple data structure having low constant factor. Using the algorithm, shortest paths from the destination to every vertex are calculated and the values in the data structure are preserved to be utilized in Step II. In the algorithm presented in [19], multiple arrays were being used like color (to keep track of the status of the nodes i.e. traversed, half traversed or not traversed), distance (each element maintain the shortest distance from source node), type (each element contains one of the values of the type i.e. CULPRIT, Skip or normal), parent (each element is having the name of the parent node). Out of these only array of parent (Π) and array of distances (Δ) are required to maintain values.

2.2 Step II—Finding Alternate Best Route at Every Vertex

Step I calculates shortest paths from every vertex to destination vertex including the original source. This step is to calculate the best path considering link failures at every point by exploring vertices in BFS fashion. On every node, it explores all edges, using equ (i) and equ (ii). Let exploring node is $\boldsymbol{u}$, node being explored is $\boldsymbol{v}$ and $\boldsymbol{minEdge1}$ and $\boldsymbol{minEdge2}$ are edges making best and second best paths from $\boldsymbol{u}$.

$$\boldsymbol{minEdge1} = \left[\forall_{v \in \boldsymbol{adj}[u]} \left(\Delta'_{\boldsymbol{u}} + \boldsymbol{e_{u,v}} + \Delta_v\right)\right]_{\boldsymbol{min}} \text{.........................equ (i)}$$

$$\boldsymbol{minEdge2} = \left[\forall_{v \in \boldsymbol{adj}[u]/\boldsymbol{minEdge1}} \left(\Delta'_{\boldsymbol{u}} + \boldsymbol{e_{u,v}} + \Delta_v\right)\right]_{\boldsymbol{min}} \text{............equ (ii)}$$

Where $\Delta'_{\boldsymbol{u}}$ is distance from source to u, $\boldsymbol{e_{u,v}}$ is cost of the edge between $\boldsymbol{u}$ and $\boldsymbol{v}$, and Δ_v is distance from destination to $\boldsymbol{v}$ (calculated in Step I)

The resultant tree, having embedded backups will be stored in two arrays of **Min**s—**Min**$^{\mathbf{1}}$ for shortest path tree edges and **Min**$^{\mathbf{2}}$ for alternate shortest path tree edges.

The algorithm is as follows:

```
Initialize all vertices other than source,s

Initialize s

Insert s into Queue

WHILE there is any element in the queue
   extract the first element from the queue and store in u
   Find the node making most minimum distance from s
   Store u against the most minimum node.
   Find the node making 2nd minimum distance from s
   Store u against the 2nd minimum node.
   Mark u as 'traversed'
 End WHILE
```

3 Working Example

In order to show and explain the algorithm a graph with following properties is being used as illustrated in figure 1:

- Graph type : ***2-edge connected, Undirected, Weighted***
- Edges weight type: ***Non-unique Positive weights***
- Total Vertices: ***11***
- Total Edges: ***22***
- Source : ***A***
- Destination : ***K***

3.1 Step I

Shortest path from **K** to every other vertex is calculated and its result along with the final values are kept in respective arrays are shown in figure 2.

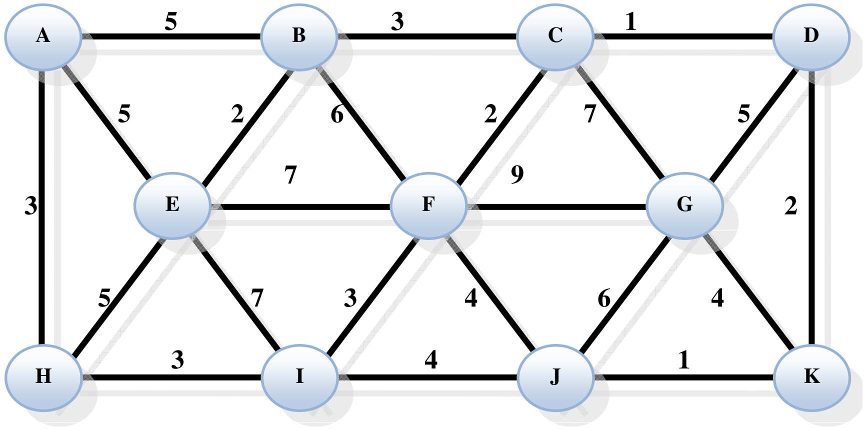


Fig. 1. Graph with 11 vertices and 22 edges

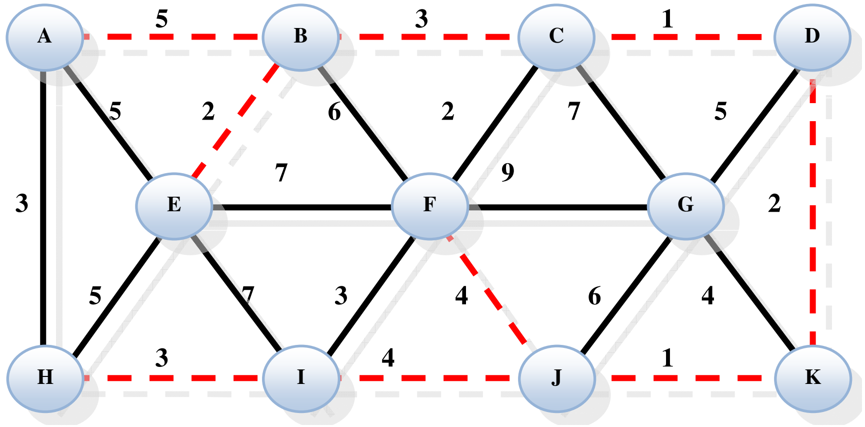


Fig. 2. Red dashed lines in the graph are showing the shortest path from K (destination) to rest of the vertices Result of Step I

3.2 Step II

This step starts with the source vertex i.e. **A**, explores all its adjacent vertices i.e. **B, E** and **H,** one by one. All the adjacent nodes are explored one by one and out of these, nodes making minimum and second minimum distance from source are calculated, marked and stored. Once all adjacent nodes have been explored and enqueued then value minimum is stored in **Min**^{1} and value of 2nd minimum is stored in **Min**^{2} and **A**'s status is updated to 'traversed'. New vertex is then extracted from the queue and the process continues until (**Q**) has any vertex in it. For detail explanation please refer to table 2.

Table 1. The array representation of the tree of step I. Parent of each node is stored as the value of that node. Second row of the table is showing the distances that are calculated during execution of the shortest path algorithm. Both of these arrays are going to be used in step II.

Node	A	B	C	D	E	F	G	H	I	J	K
Π	B	C	D	K	B	J	K	I	J	K	Φ
Δ	11	6	3	2	8	5	4	8	5	1	0

Table 2. ***u*** is the vertex whose adjacent vertices are being explored; ***v*** is one of the adjacent vertices. Minimum and second minimum are being calculated and the remaining distance is being shown along with them (written in brackets) and the distance travelled is being shown in first column along with vertex.

U(distance from ‘s’)	Adjacent Vertices	Min^1_u(dist. to ‘t’)	Min^2_u(dist. to ‘t’)	Queue
A(0)	B			
	E			
	H	B(11)	H(11)	BEH
B(5)	C			
	E			
	F	C(6)	E(10)	EHCF
E(5)	B			
	F			
	H			
	I	B(8)	F(12)	HCEFI
H(3)	E			
	I			
	A	I(8)	E(13)	CEFI
C(8)	F			
	G			
	D	D(3)	F(7)	EFIGD
F(10)	B			
	E			
	I			
	J			
	G			
	C	J(5)	C(5)	IGDJG
I(6)	E			
	H			
	F			
	J	J(5)	F(8)	GDJ
G(15)	C			
	F			
	J			
	K			
	D	K(4)	J(7)	DJK
D(9)	G			
	K	K(2)	G(9)	JK
J(10)	G			
	K	K(1)	G(10)	K

4 Discussion and Analysis

This algorithm provides link failure recovery mechanism embedded in the shortest path so that the execution of the traversal on path can be diverted to alternate paths if some hurdle appears on the best path. Along with the embedded backs it also provides the information regarding the remaining length of the path at every point that may be helpful in many applications like games. In case of link failure, not only the traversal can be continued without any delay but at the same time, the time required on all other available paths can also be seen, which may be advantageous in some future works. For details see fig 3.

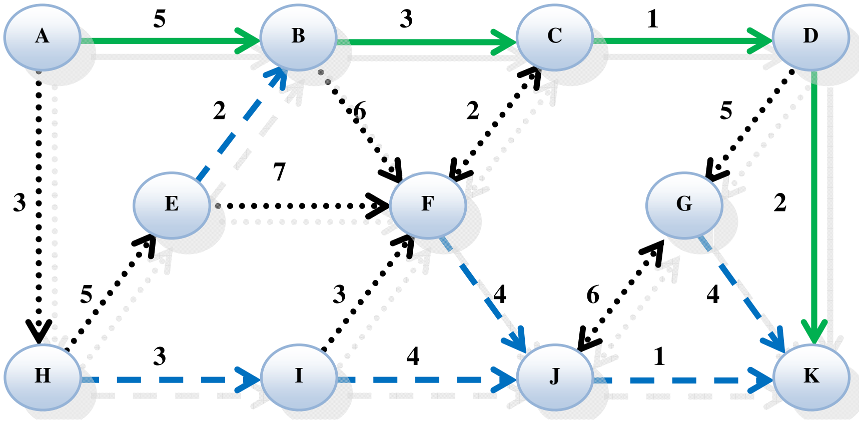


Fig. 3. Green solid lines presents shortest path from A to K. Blue dashed lines presents the best path from that vertex to destination while black doted lines presents 2nd best path from the vertex.

This algorithm calculated and embeds the backups in the time that any good shortest path algorithm require for the calculation of shortest path tree. Its complexity shows that no extra time is required to calculate and embed the backup links with the shortest path tree.

In order to maintain backup paths, algorithm needs extra memory for storage purpose which is inevitable. Algorithm itself is very simple and at the same time it is using very simple data structure. The use of simple data structure reduces not only the implementation cost but effort too. It is easy to understand and easy to implement. It is straight forward as both steps (shortest path algorithm and BFS like search) are very easy to implement and do not need high proficiency in algorithms for understanding.

4.1 Time Complexity

The algorithm is designed in two steps so the total running time or asymptotic complexity would be the sum of the running times of the two individual steps. In step-I a shortest path is being calculated using the shortest path algorithms presented in [20][19]. The complexity of the algorithm is $O\ (E\ +\ V\ log\ V)$.

In second step a BFS like traversal, explores all neighbours of all the nodes one by one. While exploring neighbors of one node it calculates nodes making the minimum and second minimum remaining distances and store them as backup path. it is a simple BFS algorithm and rest calculation are $O(1)$ time so the total time required for this step is $O(V\ +\ E)$.

Adding the two and simplifying the expression we can get over all asymptotic running time of the algorithm that is $O(E\ +\ V\ log\ V)$.

5 Conclusion

This is an efficient algorithm. It calculates and embeds shortest path as well as backup shortest paths in one tree in the time that is required to calculate shortest path only. It has the capability of overcoming multiple link failures at multiple locations but one at one spot. So we can say that it is providing single link failure recovery at any point. Not only theoretically but also practically it is suitable in applications like transportations networks and robot navigation. It does not have high constant factor which normally slows down the speed of algorithms when it comes to practical situation. It is more efficient than the existing algorithms that are discussed in introduction. Unlike Dijkstra's (Fibonacci heap implementation) it is not using any complex data structure that makes it easy to implement and easy to understand.

References

[1] Zhang, B., Zhang, J., Qi, L.: The shortest path improvement problems under Hamming distance. Journal of Combinatorial Optimization 12, 351–361 (2006)

[2] Safdar, S., Hassan, M.F.: Moving Towards Two Dimensional Passwords. In: 4th International Symposium on Information Technology, pp. 891–896 (2010)

[3] Safdar, S., Hassan, M.F.: Framework for Alternate Execution of workflows under threat. In: 2nd International Conference on Communication Software and Networks, Singapore (2010)

[4] Akbar, R., Hassan, M.F.: Limitations and Measures in Outsourcing Projects to Geographically Distributed Offshore Teams. In: 4th International Symposium on Information Technology, KL, Malaysia, pp. 1581–1585 (2010)

[5] Akbar, R., Hassan, M.F.: A Collaborative–Interaction Approach of Software Project Development–An Extension to Agile Base Methodologies. In: 4th International Symposium on Information Technology, KL, Malaysia, pp. 133–138 (2010)

[6] Qureshi, M.A., Maqbool, O.: The Complexity of Teaching: Computability and Complexity. In: International Conference on Teaching and Learning 2007, INTI International University College at Putrajaya, Malaysia (2007)

[7] Qureshi, M.A., Maqbool, O.: The Complexity of Teaching: Computability and Complexity. INTI Journal Special Issue on Teaching and Learnning 1, 171–182 (2007)

[8] Dijkstra, E.W.: A note on two problems in connection with graphs. Numerische Mathematik 1, 269–271 (1959)

[9] Hershberger, J., Suri, S., Bhosle, A.M.: On the Difficulty of Some Shortest Path Problems. ACM Transactions on Algorithms 3, 1–15 (2007)

[10] Pettie, S., Ramachandran, V., Sridhar, S.: Experimental evaluation of a new shortest path algorithm. In: Mount, D.M., Stein, C. (eds.) ALENEX 2002. LNCS, vol. 2409, pp. 126–142. Springer, Heidelberg (2002)

[11] Thorup, M.: Floats, Integers, and Single Source Shortest Paths. Journal of Algorithms 35, 189–201 (2000)

[12] Thorup, M.: Floats, Integers, and Single Source Shortest Paths *. In: Meinel, C., Morvan, M. (eds.) STACS 1998. LNCS, vol. 1373, Springer, Heidelberg (1998)

[13] Raman, R.: Recent Results on the Single-Source Shortest Paths Problem. SIGACT News 28, 81–87 (1997)

[14] Bhosle, A.M., Gonzalez, T.F.: Efficient Algorithms for Single Link Failure Recovery and Its Application to ATM Networks. In: 15th IASTED Intl. Conf. on PDCS, pp. 87–92 (2003)

[15] Bhosle, A.M., Gonzalez, T.F.: Distributed Algorithms for Computing Alternate Paths Avoiding Failed Nodes and Links. eprint arXiv:0811.1301, vol. abs/0811.1 (2008)

[16] Li, Y., Nie, Z., Zhou, X.: Finding the Optimal Shortest Path Tree with Respect to Single Link Failure Recovery. In: Fourth International Conference on Networked Computing and Advanced Information Management, pp. 412–415. IEEE, Los Alamitos (2008)

[17] Qureshi, M.A., Hassan, M.F., Safdar, S., Akbar, R., Sammi, R.: Shortest Path Algorithm With Pre-calculated Single Link Failure Recovery for Non-Negative Weighted Undirected Graphs. In: International Conference on Information and Emerging Technologies (ICIET), pp. 1–5 (2010)

[18] Qureshi, M.A., Hassan, M.F., Safdar, S., Akbar, R.: A Near Linear Shortest Path Algorithm for Weighted Undirected Graphs. In: 2011 IEEE Symposium on Computers & Informatics, KL, Malaysia (2011) (accepted)

[19] Qureshi, M.A., Hassan, M.F., Safdar, S., Akbar, R.: Two Phase Shortest Path Algorithm for Non-negative Weighted Undirected Graphs. In: IEEE 2010 Second International Conference on Communication Software and Networks, pp. 223–227 (2010)

[20] Qureshi, M.A., Hassan, M.F.: Improvements over Two Phase Shortest Path Algorithm. In: 4th International Symposium on Information Technology, ITSIM 2010 (2010)

Managing Differentiated Services in Upstream EPON

Nurul Asyikin Mohamed Radzi[1], Norashidah Md. Din[1],
Mohammed Hayder Al-Mansoori[2], and Mohd. Shahmi Abdul Majid[1]

[1] Centre for Communications Service Convergence Technologies,
Dept. of Electronics and Communication Engineering,
College of Engineering, Universiti Tenaga Nasional, 43009 Kajang, Malaysia
{Asyikin,Norashidah,mshahmi}@uniten.edu.my
[2] Faculty of Engineering, Sohar University, PO Box 44, Sohar, PCI 311, Oman
mmansoori@soharuni.edu.om

Abstract. In order to enhance the bandwidth utilization while maintaining the required quality of service (QoS), an efficient dynamic bandwidth allocation (DBA) algorithm that supports global priority has been proposed in upstream Ethernet passive optical network (EPON). The algorithm aims to allocate bandwidth to different traffic classes according to the priority as a whole. The simulation is done using MATLAB by comparing the proposed algorithm with a DBA algorithm found in the literature. The proposed algorithm shows an improvement as high as 11.78% in terms of bandwidth utilization and reduces expedited forwarding (EF), assured forwarding (AF) and best effort (BE) delay as high as 25.22%, 18.96% and 21.34% respectively compared to Min's DBA.

Keywords: EPON, DBA, global priority, differentiated services.

1 Introduction

Ethernet passive optical network (EPON) has a tree topology where it consists of an optical line terminal (OLT) and multiple optical network units (ONU)s that is connected to each other by an optical splitter. In the downstream traffic transmission, frames are broadcasted by an OLT to every ONU in the system. However, each ONU will selectively receive frames that are meant for them. In contrast, for upstream traffic transmission, multiple ONUs sent the frames to the OLT at the same time, and problem occurs since all these frames have to share the same fiber from the splitter to the OLT.

Therefore, the IEEE 802.3ah task force has developed a protocol that is used to distribute the bandwidth using time division multiplexing (TDM) [1, 2]. The protocol is called as multipoint control protocol (MPCP). MPCP provides only a framework for various dynamic bandwidth allocation (DBA) algorithms. It does not provide DBA algorithm that is used to adapt the network capacity to traffic conditions [3, 4].

The two types of control messages in MPCP are GATE and REPORT. GATE is sent by the OLT to notify ONU of the assigned timeslot. REPORT is sent by the ONUs giving the information of transmission start time and queue occupancy.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 672–679, 2011.

1.1 Related Work

Different types of traffic need different types of priority. For example, real time traffic such as voice traffic requires less delay and jitter as compared to video and data traffics, so in order to enhance the overall performance of an EPON system, quality of service (QoS) needs to be realized [5]. In a whole, we can differentiate the DBA algorithms in the literature currently to two different categories; some that support QoS and others that do not support QoS.

Among the most referenced DBA algorithm that do not support QoS is Interleaved Polling with Adaptive Cycle Time (IPACT) [6, 7]. IPACT can be divided into several approaches such as fixed, gated and limited. It considers each ONU as one without differentiating the traffics as according to their priority.

Another example of DBA algorithm that does not support QoS is Weighted DBA [8]. In Weighted DBA, traffics are divided according to the arrival of packets. Traffics that arrived before sending the REPORT message are being granted first regardless whether it is real time traffic or not. This causes a term that is called as light load punishment, where the traffic that is delay sensitive needs to wait for other traffics that is not delay sensitive to be granted first before their turn arrive.

An example of DBA algorithm that supports QoS is Broadcast Polling (BP) algorithm [9]. With BP algorithm, ONUs are divided into three different categories and will be served according to the priority. The details on how they categorized the ONUs are not stated in the paper. However, we cannot have only voice traffic inside a particular ONU and only video traffic in the other ONUs. Since we are living in a converged network, each ONU must be divided into different queues and each queue must be treated according to the QoS.

Thus, Min's DBA [10] overcomes the problem by having hierarchical scheduling where the OLT divides the bandwidth to the ONUs and each ONU further divides the traffic into three different queues inside that particular ONU to realize the QoS. However with this way, QoS is realized locally, which is inside the ONU itself only.

Therefore, we proposed a DBA algorithm that we called as Efficient DBA with Global Priority (EDBAGP) that supports the QoS globally. We compare the algorithm with Min's DBA algorithm and it shows an enhancement in the overall performance, where we improve the bandwidth utilization as high as 11.78% and reduce the packet delay as high as 25.22%.

This paper is organized as follows. We explained about our DBA algorithm in Section 2, the result in Section 3 and the conclusion in Section 4.

2 Methodology

In EDBAGP, we have three buffers inside each ONU and inside the OLT. The three buffers correspond to the three different types of priorities; highest priority, medium priority and lowest priority.

Expedited forwarding (EF) bandwidth is considered as the highest priority because it supports voice traffic that requires bounded end-to-end delay and jitter specifications. Assured forwarding (AF) bandwidth that supports video traffic is considered as the medium priority because it is not delay sensitive although it requires

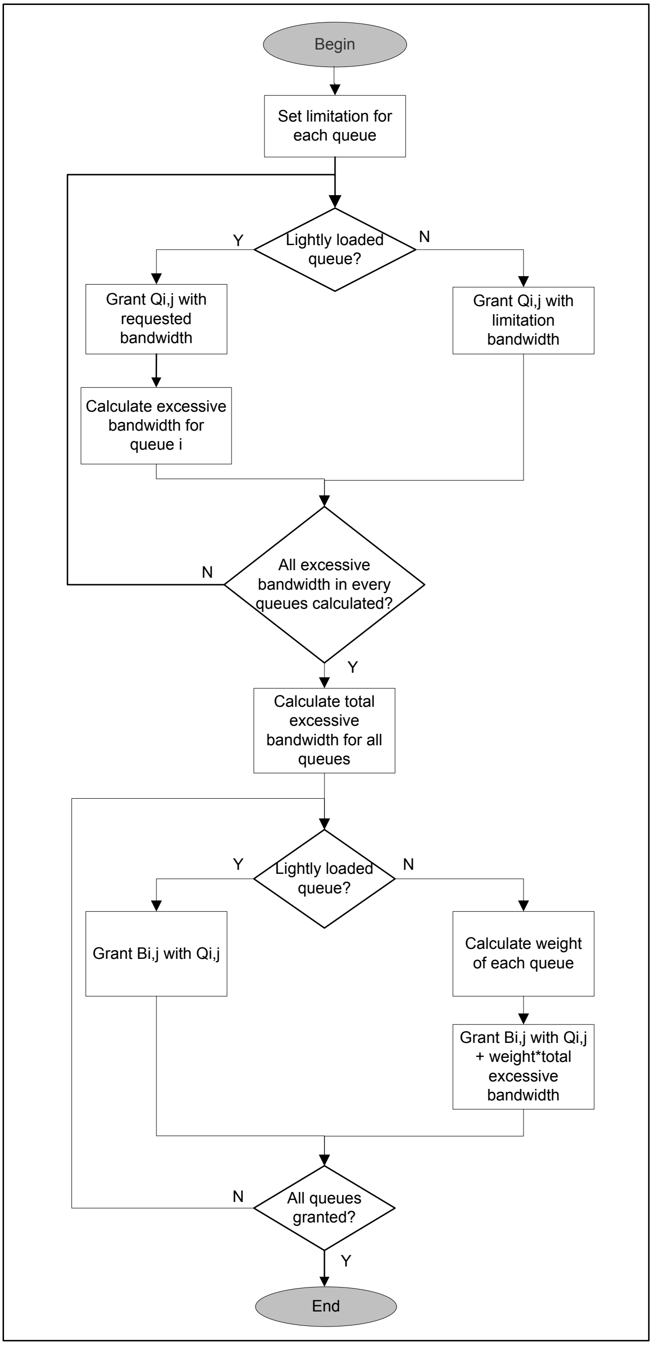

Begin
Set limitation for each queue
Lightly loaded queue?
Y
N
Grant Qi,j with requested bandwidth
Grant Qi,j with limitation bandwidth
Calculate excessive bandwidth for queue i
All excessive bandwidth in every queues calculated?
N
Y
Calculate total excessive bandwidth for all queues
Lightly loaded queue?
Y
N
Grant Bi,j with Qi,j
Calculate weight of each queue
Grant Bi,j with Qi,j + weight*total excessive bandwidth
All queues granted?
N
Y
End

Fig. 1. Flowchart for bandwidth allocation in EDBAGP

bandwidth guarantees. Finally, best effort (BE) bandwidth that supports data traffic is considered as the lowest priority as it is not sensitive to end-to-end delay or jitter.

EDBAGP is a hierarchical scheduling where the top level scheduler is used to calculate the excessive bandwidth and the lower level is to distribute that excessive bandwidth and it can be better described in the flowchart in Figure 1.

We first set a limitation bandwidth for each queue inside ONU. For this paper, we set the limitation bandwidth for EF to be 3100 bytes, AF and BE to be 6200 bytes each. Then we check if the requested bandwidth for a particular queue is less than the limitation bandwidth that has been set. If yes, we call it as lightly loaded queues, else we call is as highly loaded queues.

For lightly loaded queue, we grant the queue *i* for ONU *j*, $Q_{i,j}$ with as high bandwidth as requested. Then we calculate the excessive bandwidth, $B_{ei,j}$ of queue *i* in *j* ONU which we obtain by using the formula in Equation 1.

$$B_{ei,j} = S_i - R_{i,j} \ . \tag{1}$$

where S_i is the limitation for queue *i* and $R_{i,j}$ is the requested bandwidth for queue *i* in *j* traffic.

For highly loaded queue, we grant $Q_{i,j}$ with the limitation bandwidth.

After all queues have been granted in the first stage, we calculate the total excessive bandwidth in lightly loaded queues. Then, we grant lightly loaded queues with $Q_{i,j}$.

For highly loaded queue, we calculate the weight of each queue using Equation 2.

$$B_{excess} = \frac{w_{i,j}}{\sum_{highly\ loaded} w_{i,j}} B_{e,total} \ . \tag{2}$$

where $w_{i,j}$ is the weight of queue *i* in ONU *j*. Then, we grant the highly loaded queue with the summation of $Q_{i,j}$ and B_{excess}.

3 Performance Evaluation

We compare our proposed DBA algorithm, EDBAGP with Min's DBA in order to verify our analysis and demonstrate its performance. We used the same parameters for both algorithms and the parameters can be shown in Table 1. The comparison is done by using simulation and we simulate both algorithms using MATLAB.

The simulation result shows that both algorithms can perform the bandwidth allocation efficiently. However, EDBAGP shows better performance than Min's DBA as can be seen in Figure 2.

Table 1. Simulation parameters

Parameters	Values
Number of ONUs	16
Upstream bandwidth	1Gbps
Maximum cycle time	2ms
Guard time	5µs
Limitation EF bandwidth	3100 bytes
Limitation AF bandwidth	6200 bytes
Limitation BE bandwidth	6200 bytes

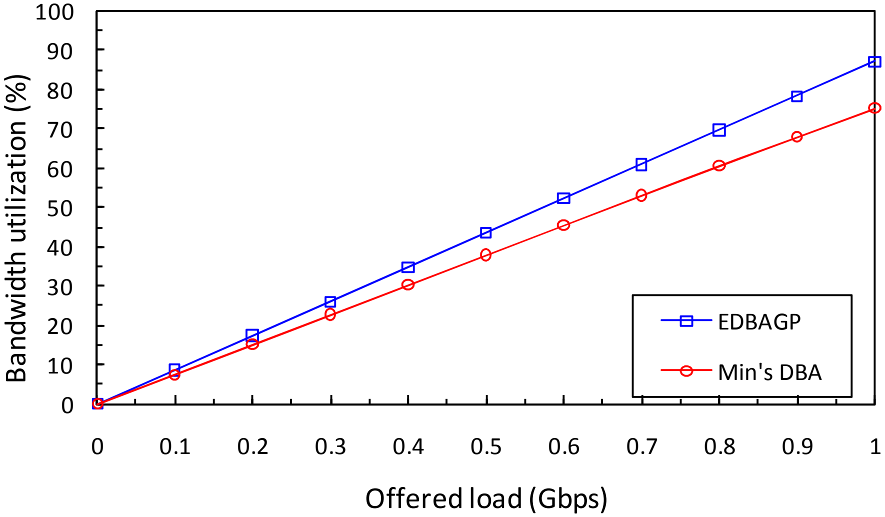


Fig. 2. Bandwidth utilization versus offered load between EDBAGP and Min's DBA algorithm

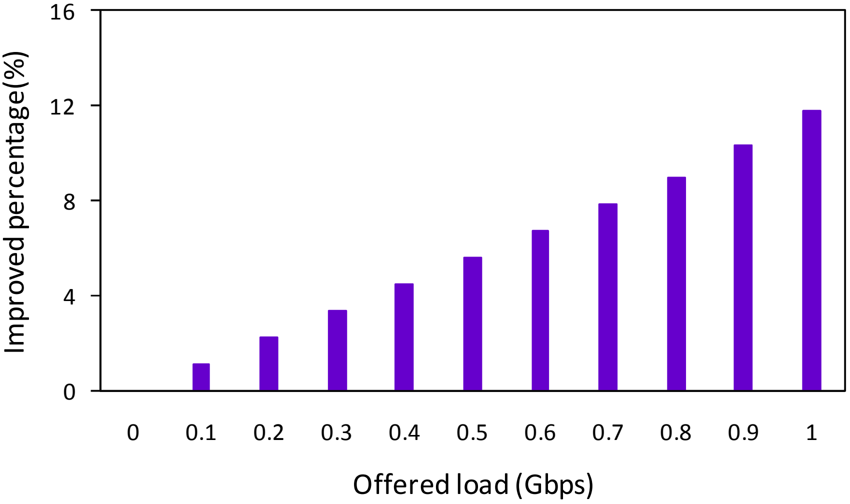


Fig. 3. Improved percentage versus offered load between EDBAGP and Min's DBA algorithm

In Figure 2, we can observe that the bandwidth utilization increases linearly as the offered load increases. EDBAGP can reach as high as 87% bandwidth utilization as the offered load is maximum, whereas Min's DBA reaches only up to 75.22%. The reason is due to the way bandwidth is being allocated in both algorithms. EDBAGP has a global priority, meaning that it does not set any limitation as of how much bandwidth can be granted in a particular ONU as what have been done in Min's DBA.

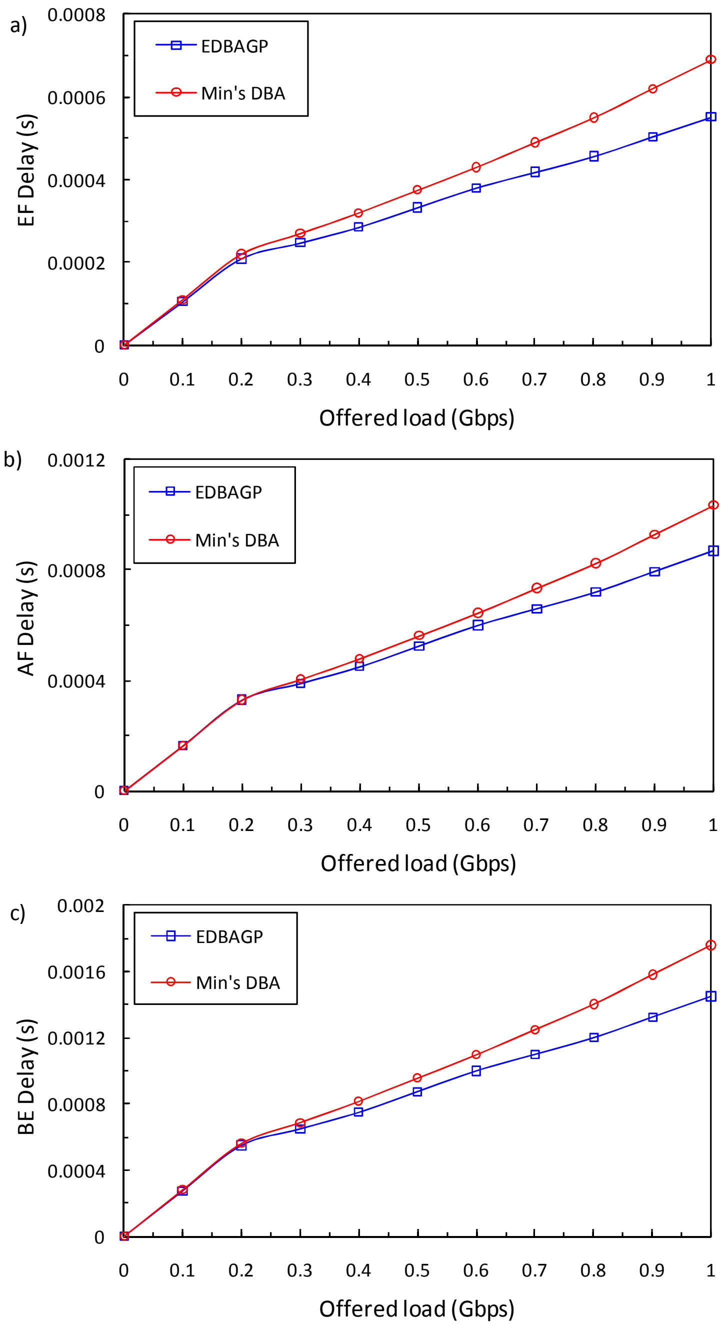


Fig. 4. a)EF b)AF and c)BE delay versus offered load for EDBAGP and Min's DBA

Therefore bandwidth allocation is 100% flexible in EDBAGP. It does not set limitation to EF bandwidth and it makes full use of the excessive bandwidth. Due to this reason, EDBGP has better performance in terms of bandwidth utilization.

We have also calculated the improved percentage of EDBAGP as compared to Min's DBA. The result is shown in Figure 3. Again, we vary the offered load from 0 to 1Gbps. The improved percentage increases linearly until it gets as high as 11.78% as the offered load is 1Gbs.

Figure 4 shows the packet delay versus offered load for the three types of traffic; EF, AF and BE between EDBAGP and Min's DBA. In all three different types of traffic, packet delays are lower than EDBAGP as compared to Min's DBA although the percentage different differs from one traffic to another.

In Figure 4a that shows EF packet delay, we can observe that as the offered load increases to 0.1Gbps, both algorithms have the same delay performance, which are 0.11ms. As the offered load increases beyond 0.1Gbps, EDBAGP starts to show better performance than Min's DBA. As it increases to 0.2Gbps, EDBAGP starts to have 0.2ms of delay compared to 0.21ms in Min's DBA. The EF delay for EDBAGP then increases to 0.55ms as the offered load is 1Gbps, whereas Min's DBA gets all the way to 0.69ms.

In Figure 4b and 4c, the delay performance for AF and BE traffics for both DBA algorithms are the same as the offered load increases to 0.2Gbps. For AF traffic, the delay shows as high as 0.33ms and for BE as high as 0.55ms. EDBAGP continues to increase as high as 0.87ms as the offered load is 1Gbps for AF traffic, whereas Min's DBA gets to as high as 1.04ms. In BE traffic, as the offered load is 1Gbps, EDBAGP reaches only as high as 1.45ms whereas Min's DBA reaches all the way to 1.76ms.

EDBAGP has lower delay than Min's DBA because it grants bandwidth according to global priority. EDBAGP grants bandwidth to lightly loaded EF traffics first, then highly loaded EF traffics, followed by lightly loaded AF traffics and so on. As differ from Min's DBA, it grants EF traffic in lightly loaded ONUs first, then AF and BE traffic in lightly loaded ONUs before granting the EF traffic in highly loaded ONUs. Thus, delay is lower in EDBAGP because it reduces the light load punishment effect.

4 Conclusion

This paper studies a DBA algorithm with differentiated services that support global priority in EPON that we called as EDBAGP. We propose a hierarchical scheduling, where the top level scheduler allocates the bandwidth up until the limitation bandwidth and the lower level scheduler allocates the excessive bandwidth to highly loaded queue in the system. We compare EDBAGP with DBA algorithm that supports differentiated services in local priority, Min's DBA. The simulation results show that EDBAGP improves as high as 11.78% in terms of bandwidth utilization compared to Min's DBA. EDBAGP reduces EF, AF and BE delay as high as 25.22%, 18.96% and 21.34% respectively compared to Min's DBA.

References

1. Ethernet in the First Mile Task Force IEEE Std 802.3ah (2004), `http://www.ieee082.org/3/efm`
2. Ngo, M.T., Gravey, A., Bhadauria, D.: Controlling QoS in EPON-Based FTTX Access Networks. In: Telecommun System, pp. 1–15. Springer, Netherlands (2010)
3. Merayo, N., Durán, R.J., Fernández, P., Lorenzo, R.M., Miguel, I.D., Abril, E.J.: EPON Bandwidth Allocation Algorithm Based on Automatic Weight Adaptation to Provide Client and Service Differentiation. In: Photonic Network Communications, vol. 17, pp. 119–128. Springer, Netherlands (2009)
4. Choi, S.I., Part, J.: SLA-Aware Dynamic Bandwidth Allocation for QoS in EPONs. Journal of Optical Communication Network 2, 773–781 (2010)
5. Radzi, N., Din, N., Al-Mansoori, M., Mustafa, I., Sadon, S.: Intelligent Dynamic Bandwidth Allocation Algorithm in Upstream EPONs. Journal of Optical Communication Networking 2, 148–158 (2010)
6. Kramer, G., Mukherjee, B., Pesavento, G.: Interleaved Polling with Adaptive Cycle Time (IPACT): A Dynamic Bandwidth Distribution Scheme in an Optical Access Network. In: Photonic Network Communication, pp. 89–107 (2002)
7. Kramer, G., Mukherjee, B., Pesavento, G.: IPACT A Dynamic Protocol for an Ethernet PON (EPON). IEEE Communication Magazine 40, 74–80 (2002)
8. Bai, X., Shami, A., Assi, C.: On the Fairness of Dynamic Bandwidth Allocation Schemes in Ethernet Passive Optical Network. Computer Communications 29, 212–2135 (2006)
9. Xiong, H., Cao, M.: Broadcast Polling-An Uplink Access Scheme for the Ethernet Passive Optical Network. Journal of Optical Networking 3, 728–735 (2004)
10. Min, L., Xiaomei, F., Yu, C., Fulei, D.: New Dynamic Bandwidth Allocation Algorithm for Ethernet PON. In: 8th International Conference on Electronic Measurement and Instruments, vol. 3, pp. 224–227 (2007)

Real-Time Volume Shadow Using Stencil Buffer

Hoshang Kolivand[1], Mohd Shahrizal Sunar[1], Yousef Farhang[2,3], and Haniyeh Fattahi[3]

[1] UTM ViCubelab, Faculty of Computer Science and Information Systems,
Department of Computer Graphics and Multimedia, Universiti Teknologi Malaysia,
81310 Skudai Johor, Malaysia
[2] Department of Computer Science, Islamic Azad University Khoy Branch,
Khoy, Iran
[3] Faculty of Computer Science and Information System, Universiti Teknologi Malaysia,
81310 Skudai Johor, Malaysia

Abstract. Two of the best methods to recognize silhouette to create real-time volume shadow in virtual environment are described in this paper. Volume shadow algorithm is implemented for virtual environment with moveable illuminated light source. Triangular method and the Visible-non visible method are introduced. The recent traditional silhouette detection and implementation techniques used in volume shadow algorithm are improved. With introduce flowchart of both algorithms, the last volume shadow algorithm using stencil buffer is rewritten. A very simple pseudo code to create volume shadow is proposed. These techniques are poised to bring realism into commercial games. It may be use in virtual reality applications.

Keywords: Real-time, Volume shadow, Silhouette detection, Stencil buffer.

1 Introduction

To have a realistic environment, shadow is the most important effect that reveals more information about the distance between objects in the scene. It is the major factor of 3-D graphics for virtual environment but unfortunately is difficult to be implemented in display environment especially in real-time games. In computer games, shadows give the gamers feelings that could trigger an imagination that they are playing in the real world hence provides maximum pleasure. A game with lack of shadow cannot be attractive especially in this century that gamers' imagination requests more realistic situation when they are watching cartoons or playing games.

There are many algorithms to create shadow but volume shadows have a great achievements in game makers. Although volumes shadow is considered as established in gaming industry, they have two expensive phases. One of them is update of volume rendering passes and the other one is silhouette detection.

Recognizing the outline of object can increase the speed of algorithm. To find the outline of object, silhouette detection is essential because it can reduce the cost of implementation and it is the main item to improve an algorithm. Silhouette detection is an important phase in all visual effects. It is mentionable that the bases of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 680–687, 2011.

silhouettes are visual and view point. To generate shadow, silhouette detection plays a crucial role to detect the boundary of occluder.

The idea of shadow volume was first introduced in 1977 when Frank Crow [1] published his ray-casting based shadow volume algorithm. His method explicitly clips shadow geometry to view the frustum. In 1991 Heidmann published a paper based on volume shadow using stencil buffer which is even today the main shadow volume algorithm [2]. Stencil shadows belong to the group of volumetric shadow algorithms, as the shadowed volume in the scene is explicit in the algorithm. Let's take a look at how the original stencil shadow volume technique works.

The previous algorithms include rays that are traced from infinity towards the eye [3], [4]. Shadow Maps suggested by Fernando et al [5] which are an extension to the traditional shadow mapping technique. In year 2002 Lengyel propose a hybrid algorithm that uses faster Z-pass rendering [6].

Benichou and Elber[7] used a method to compute parallel silhouette edges of polyhedral object based on Gaussian sphere, for the normal vectors which are located on the view direction mapped onto a Gaussian sphere. Matt Olson and Hao Zhang [8], in 2008 work on tangent-space and they focused on tangential distance of objects to be used in polygon mesh silhouette detection. In 2010, Matt Olson [9-10], designed a site that lists all related papers.

2 Silhouette Algorithms

Silhouette detection is a function like f: $R^3 \rightarrow R^2$. Silhouettes have important role to create shadow on shadow receiver. To create shadow, projection of outline of object is enough to generate shadow as a result the cost of projection will be decreased. The most expensive part of shadow volume algorithm is identification silhouette of occluder.

Silhouettes have the most important role to recognize and project shape onto the shadow receiver. To create shadow, projection of silhouette of occluder is enough to generate a shadow of the whole object and as a result the cost of projection will be low. A silhouette edge of polygon is the edges that belong to two neighborhood planes where normal vector of one of them is towards the light and normal vector of the other plane is away from the light. If the volume shadow is desired point by point, it is too difficult to execute the program. It needs a lot of calculation and it takes a substantial CPU time for rendering. To improve this technique we should recognize the contour edges or silhouettes of object and implement the algorithm just for silhouettes. Using silhouette edges of the occluder to generate a volume shadow could optimized the process because the amount of memory is decreased, therefore, rendering will be done faster. The silhouette should be recalculated when position of the light source changes or the occluder moves.

There are many methods to recognize outline of object. Most of them are expensive but our approach is to introduce an algorithm that is not expensive like past algorithms. In stencil shadow algorithm it is needed to divide silhouette face of occluder to triangle meshes. It means that each edge should shared by just two triangles. To determine which edge is silhouette, it is needed to recognize which aged is shared between a face toward the light source and a face away to the light source.

2.1 Triangular Algorithm

In this algorithm it is needed to divide each face of occluder into triangle meshes. It means that each edge should be shared by just two triangles. To determine which edge is silhouette, it is needed to know which edge is shared between faces towards the light source and faces away of the light source.

Here the triangular algorithm is introduced:

$$S=\sum_{i}^{P}\sum_{j}^{N_i} iif\,(v_{ij}\in S[u_{ij}], Delete\ v_{ij}\ from\ S, Add\ u_{ij}\ to\ S)$$

Where:

P : Number of all polygons

N_i: Edge number of ith polygon

v_{ij}: First vertex of edge ith

u_{ij}: Second vertex of edge jth

S: An array of vertices

Fig 1 illustrates how to divide one side of a box which is consisting of four triangles. To determine silhouette we need just outline of the box not all of the edges.

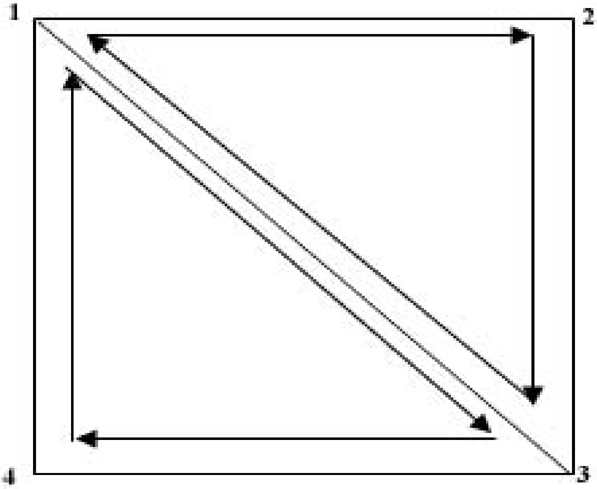


Fig. 1. Silhouette determination

2.2 Visible-Non Visible Algorithm

In this algorithm, all edges which have just one visible face are silhouette. These edges which have exactly one visible and one invisible face are silhouette, but on the contrary, each edge with two visible faces or two invisible faces is not regarded as a silhouette.

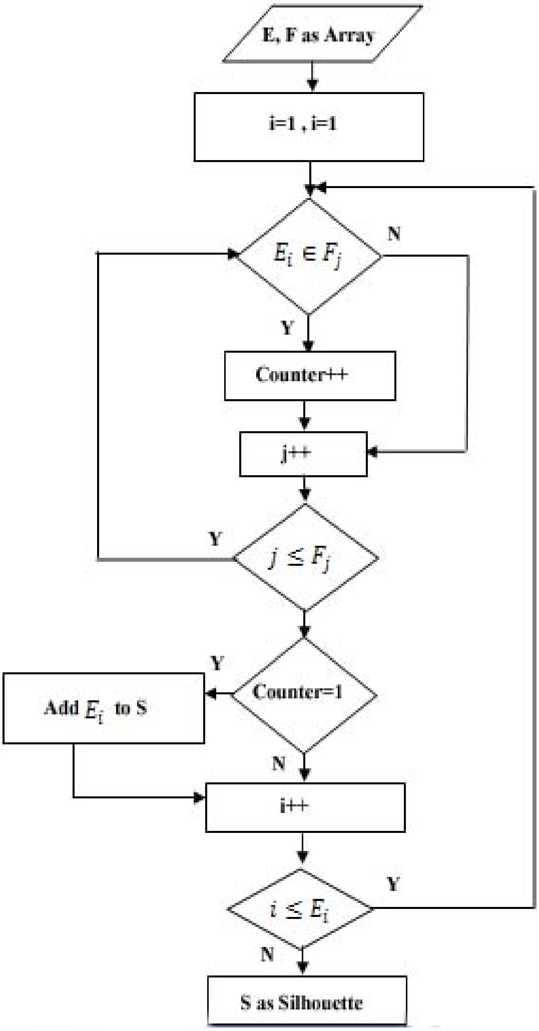


Fig. 2. Visible-non visible silhouette determination flowchart

It is important to note that silhouette determination is one of the two most expensive operations in stencil shadow volume implementation. The other is the shadow volume rendering passes to update the stencil buffer. These two steps are the phase that we intend to consider in the nearest future.

3 Stencil Buffer

Stencil buffer is a part of memory in 3-D accelerator that can control rendering of each pixel in scene. In other word the stencil buffer is the number bits in the frame buffer that is used for enabling and disabling drawing of each pixel of object in rendering, it can also be used to mask color buffer.

There is a close relationship between stencil buffer and Z-buffer. Both these buffers are neighborhood and are located in the graphics hardware memory. Z-buffer requires controlling a selected pixel of stencil value that is increased or decreased when it is needed to count the time of entering and leaving the shadow volume respectively. The stencil buffer will be increased when eye's ray enters the volume shadow by passing from front face of volume shadow and it will be decreased when ray leaves the volume shadow by passing of back face of it. It has two phases, first render the front faces; if depth test passes then stencil buffer should be increased. In the second rendering; if depth test passes stencil buffer should be decreased.

Nowadays, Opengl and DiretX support stencil buffer, so in our approach depth buffer and stencil buffer is combined together.

4 Using Visible-Non Visible Algorithm to Create Volume Shadow Using Stencil Buffer

Now if each object or part of object that is inside of truncated pyramid is in shadow and it should be dark but each object or part of object that is out of truncated pyramid is in lit and it should not be dark To generate volume shadow using stencil buffer and depth buffer together, following steps should be done:

1-Render the whole scene just with ambient.
The color buffer will be filled for all pints of object in shadow and also Z-buffer will be filled by depth value.
2-After disable Z-buffer and color buffer, render whole scene with lighting.
3-Subtract of this two depth value provides volume shadow.

After generating volume shadow, we should pass the ray from the eye to each point of object on the shadow receiver. As it mentioned before (in stencil buffer) ,if the ray pass the front face of volume shadow, increase the stencil buffer and when the ray pass the back face of volume shadow, decrease the stencil shadow. Finally, if the number of stencil buffer is not zero it is in shadow. We are going to say this as an algorithm with silhouette detection:

Here the volume shadow algorithm is introduced as a pseudo code:

```
1-Provide the equipment to have stencil buffer
2-Render the all scene without lighting
3-Enable stencil buffer
4 -Disable depth buffer to write and prevent to write in color buffer.
5- Use silhouette detection
```

```
6- For Each Plane (P)

    If DotProduct(P,R)>0
        For Each Pixel
            If (Z-test passes)
                Increase StencilBuffer
7- For Each Plane (P)

    If DotProduct(P,R)<0
        For Each Pixel
            If (Z-test passes)
                Decrease StencilBuffer
8-Render the all scene again with lighting
9-Enable to write in color buffer
10-If (the stencil buffer mod 2=0)
    The pixel is out of shadow
```

Z-pass algorithm is used when eyes are out of shadow. If eyes are located in shadow Z-fail algorithm is used.

5 Results

In each algorithm, frame per second (FPS) is the most important factor to implement. The triangular method and visible- none visible method are used to recognize silhouette. It is amazing that the visible-non visible method has higher FPS than the triangular method.

To have shadow on arbitrary object such as torus or sphere stencil buffer and Z-pass algorithm are used. In following table the FPS for using triangular algorithm and visible-non visible algorithm is compared.

According to Figure 3, 4 and results of Table 1, both proposed algorithm are convenient to implement hard shadow in virtual environment. They can be possible to use for commercial games.

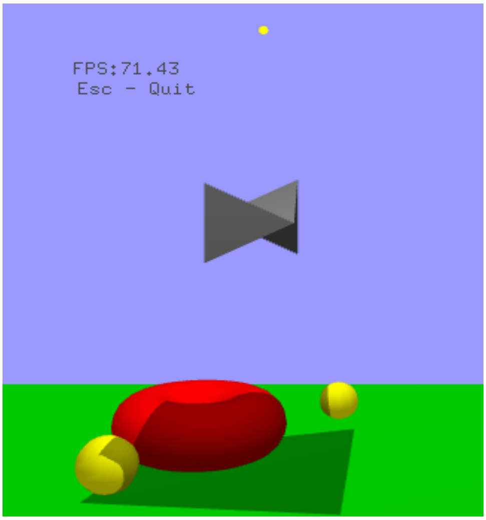
Different between FPS when project is rendered without shadow and when is rendered with volume shadow using Triangular algorithm for silhouette detection is not so much. In additional, in this case FPS is acceptable.

The FPS indicates volume shadow using Visible-non visible algorithm for silhouette detection is better than Triangular detection.

In addition, the following table illustrates the proposed algorithm for silhouette detection is appropriate for using in real-time shadow in virtual environment.

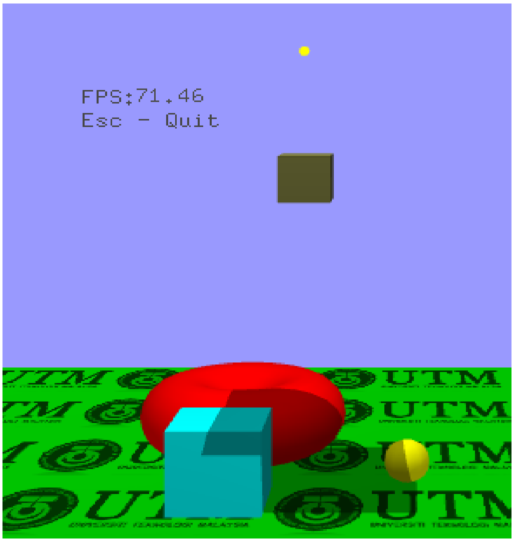
Table 1. Compare FPS

Status	*FPS*
Without shadow	71.47
Using Triangular Algorithm	71.43
Using Visible-Non visible Algorithm	71.46



FPS:71.43
Esc - Quit

Fig. 3. Volume Shadow Using Triangular Algorithm



FPS:71.46
Esc - Quit

Fig. 4. Volume Shadow Using Visible-Non Visible Algorithm

6 Conclusion

In this paper two kinds of silhouette detection were presented. To create volume shadow, triangular algorithm and visible-non visible algorithm are used. To have volume shadow on an arbitrary object stencil buffer is used. Z pass method is an appropriate tool to check if an object or part of object is located inside the volume shadow or is located in lit. Volume shadow that was difficult to understand and implement is improved.

To make easy of the volume shadow, a pseudo code is proposed and for recognizing silhouette two flowcharts are introduced. In comparison, triangular algorithm has low FPS than the visible-none visible algorithm. To create real-time shadow on arbitrary object volume shadow using stencil buffer and visible –non visible algorithm is convenient.

References

1. Crow, F.C.: Shadow Algorithms for Computer Graphics. Comp. Graph. 11, 242–247 (1977)
2. Tim, H.: Real Shadows Real Time. Iris Univ. 18, 23–31 (1991)
3. Carmack, J.: Carmack on Shadow Volumes (Personal Communication between Carmack and Kilgard), http://developer.nvidia.com/view.asp?IO=robust_shadow_volumes (referenced:10.4.2002)
4. Lokovic, T., Veach, E.: Deep Shadow Maps. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 385–392 (2000)
5. Fernado, R., Fernadez, S., Bala, K., Greenberh, D.P.: Adaptive Shadow Maps. In: Proceedings of ACM SIGGRAPH ACM Press / ACM SIGGRAPH, Computer, pp. 387–390 (2001)
6. Eric, L.: Mathematics for 3D Game Programming & Computer Graphics. Charles River Media, 437–451 (2002)
7. Benichou, F., Elber, G.: Output Sensitive Extraction of Silhouettes from Polygonal Geometry. In: Proceedings of Pacific Graphics 1999, pp. 60–69 (1999)
8. Zhang, H.: Forward Shadow Mapping. In: Proceedings of the 9th Eurographics Workshop on Rendering, pp. 131–138 (1998)
9. Olson, M., Zhang, H.: Silhouette Extraction in Hough Space. Comp. Graph. Forum 25, 273–282 (2006)
10. Olson, M.: A Survey of Silhouette Papers (2010), http://www.cs.sfu.ca/~matto/personal/sil_papers.html

Highest Response Ratio Next (HRRN) vs First Come First Served (FCFS) Scheduling Algorithm in Grid Environment

Rohaya Latip and Zulkhairi Idris

Institute for Mathematical Research (INSPEM)
University Putra Malaysia,Serdang 43400,Selangor
rohaya@fsktm.upm.edu.my

Abstract. Research on Grid scheduling nowadays, focuses in solving three problems such as finding a good algorithm, automating the process, and building a flexible, scalable, and efficient scheduling mechanism. The complexity of scheduling predicament increases linearly with the size of the Grid. Submitted jobs in Grid environment are put in queue due to the large number of jobs submission. An adequate Grid scheduling technique used to schedule these jobs and sending them to their assigned resources. This paper simulates in C programming First Come First Served (FCFS) and Highest Response Ratio Next (HRRN) Grid scheduling algorithms. A good scheduling algorithm normally shows lower value of total waiting and schedule time. Hence, HRRN was selected because of the algorithm outperform the existing gLite Grid middleware scheduling. From the simulation result proof HRRN has better performance of total waiting time due to the priority scheme policy implementation in scheduler.

Keywords: Grid computing, scheduling, simulation, Highest Response Ration Next Algorithm.

1 Introduction

The scheduling predicament, in general, has been studied broadly in many areas, especially in transportation systems, industrial operations, and system control. Today, the scheduling in a Grid computing involves much manual administrative work. Research on Grid scheduling focuses on solving three problems: finding a good schedule, automating the scheduling process, and building a flexible, scalable, and efficient scheduling mechanism.

Grid consists of geographically distributed and heterogeneous communication, computation and storage resources that may belong to different administrative domains, but can be shared among users [1].

The most significant advantage of Grid computing is collaboration of many computers (resources) in a network forming one single powerful computer to perform processing of extra ordinary tasks, which requires a great number of computing speed

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 688–693, 2011.

and processing capabilities. Grid Computing can be thought of as distributed and large-scale cluster computing and as a form of network-distributed parallel processing. The aim of the Grid is to provide the ability to manage all the computing resources across different administrative domains which cannot be completed by traditional clusters or distributed computing.

Submitted jobs in Grid environment are put in queue due to the large number of jobs submission. Therefore an efficient scheduling technique is needed to manage the jobs efficiently. Throughout the years, many scheduling techniques were researched but still there is a long waiting time to schedule these jobs and sending them to their assigned resources.

The main purpose of scheduling is to satisfy users and system demands in dealing with waiting time of a job. An efficient scheduler is when the scheduler can pass the job to the resources at a low schedule time with minimum waiting time.

Many scheduling algorithms were used to minimize the waiting time in Grid computing. One of it is First Come First Served (FCFS) which seems inefficient when dealing with large sizes jobs where small sizes jobs need to wait a long time to be process whenever there are large sizes jobs queuing in front. Hence, this paper focuses on the performance of Highest Response Ratio Next (HRRN) scheduling algorithm which promotes “aging priority scheme”. The HRRN was chosen because of its heuristic feature that deems as important to the Grid scheduling. Computer simulation was used to demonstrate this inspiring result.

2 Related Works

Schedulers are responsible for the management of jobs, such as allocating resources needed for any specific job, partitioning of jobs to schedule parallel execution of tasks, data management, event correlation, and etc. These schedulers then form a hierarchical structure, with meta-schedulers that form the root and other lower level schedulers, while providing specific scheduling capabilities that form the leaves [2].

This section gives review a set of heuristic algorithms with schedule meta-task to heterogeneous computing system. Meta-task can be defined as a collection of independent tasks with no data dependencies [3]. In static computing environment, each machine executes a single task at a time.

Heuristic algorithms are more adaptive to the Grid environment [4] where both applications and resources are highly diverse and dynamic. Heuristic algorithms will make the most realistic assumptions about a prior knowledge concerning process and system loading characteristics. It also represents the solutions to the scheduling problem which cannot give optimal answers but only require the most reasonable amount of cost and other system resources to perform their function. A number of heuristic algorithms have been introduced to schedule task on heterogeneous machine. Among them are Dispatching Rule (DR) [5], Ant Colony Optimization (ACO) [6], First Come First Served (FCFS) [7], and Shortest Process First (SPF) [8].

First Come First Served (FCFS) [7] is a heuristic of this algorithm is on the priority scheme. It promotes the first task arrived to be processed first ignoring the waiting time and data sizes. In fact, this simply implements a non-preemptive scheduling algorithm. When a process becomes ready it is added to the tail of ready queue. Processes are dispatched according to their arrival time on the ready queue. The FCFS concept scheduling is fair in the formal sense of fairness but it is unfair in the sense that long jobs make short jobs wait and unimportant jobs make important jobs wait.

FCFS is more predictable than other schemes since it offers time. FCFS scheme is not useful in scheduling interactive users because it cannot guarantee good response time. The code for FCFS scheduling is simple to write and understand. One of the major drawbacks of this scheme is that the average time is often quite long.

Research shows that DR, ACO, FCFS and SPF are not producing good schedules. In [9], the performance of these algorithms was experimented in a small Grid testbed, and was shown to outperform the existing gLite. Therefore, HRRN was selected in this paper for its efficiency.

3 The Model

The model presented in this paper is Highest Response Ratio Next (HRRN) scheduling algorithm. HRRN is on the aging priority of processes, which is computed on-demand. This algorithm corrects the weakness of the Shortest Process First (SPF), where job with a long process time will always be at the back of the queue no matter how long it waits in the queue.

The SPF algorithm is biased towards the processes with short Service Times. This keeps the longer processes waiting in the ready queue for a long time, despite of arriving in the ready queue before the short jobs. In the HRRN algorithm, aging priority is a function of not only the Service Time, but also of the time spent by the process waiting in the ready queue in a non-preemptive scheduling algorithm. Once the process obtains the control of the processor, it completes to completion.

HRRN assigns job *i* to resource *j* with the highest priority value calculated using formula $(w + s) / s$. Jobs waiting in the queue will be rearranged according to the highest priority value taking the lead and consequently followed by the lower priority value. The priority scheme take into consideration time spent waiting in the queue w and expected service time s.

The scheduler algorithms implemented in this paper are HRRN and FCFS. Both algorithms use priority ordering policy where tasks are processed according to the task at the most front slot of the queue. What differentiates the two algorithms is the way the tasks are organized in the queue. HRRN reorders the tasks according to the highest priority value computed using $(w + s) / s$ formula. Where else, FCFS puts the first arrived task in front followed consecutively by the later arrivals. Below pseudo code summarizes the explanation;

```
If there is any resource free do
    for each job $i$ queued in the scheduler's ordered list do
        Compute Priority = $(w + s) / s$
        Reorder job $i$ according to priority
        Send the scheduling decision to the user of job $i$
    end for
```

Where; w : time spent waiting for the processor; s : expected service time

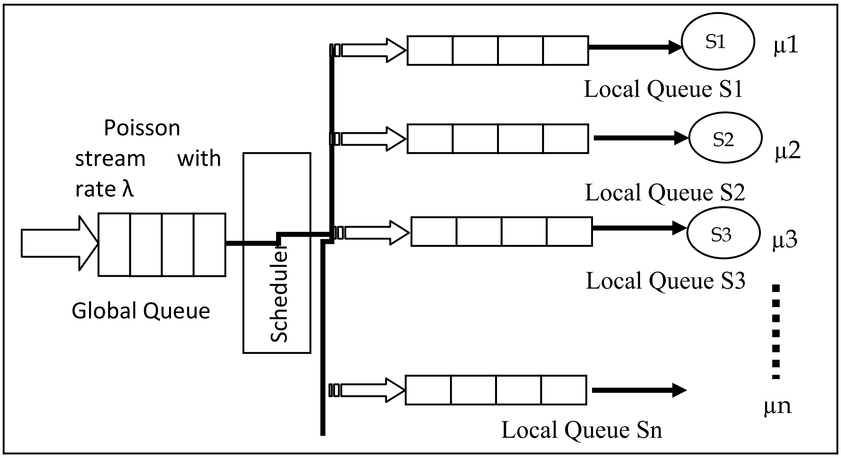


Fig. 1. Local Queue and Global Queue in HRRN

4 Implementation

HRRN was simulated in C programming language and Microsoft Visual C++ 2010 Express was used as the compiler and programming platform. There are six associated time variables for each job [10]. Fig. 2 shows the inter relationship between variables.

i) The arrival time of job *i* is a_i.

ii) The delay of job *i* in the queue is $d_i \geq 0$.

iii) The time that job *i* begins service is $b_i = a_i + d_i$.

iv) The service time of job *i* is $s_i > 0$.

v) The wait of job *i* in the service node (queue and service) is

$$w_i = d_i + s_i.$$

vi) The time that job *i* completes service (the departure time) is

$$c_i = a_i + w_i.$$

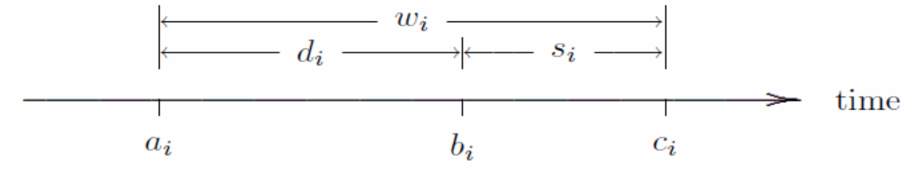


Fig. 2. Associated time variables

Based on the definition by [10], scheduling algorithms implemented in this paper was evaluated by the performance metrics as below:

i) Job waiting time : Waiting time of the job in the queue, wt_i where

$$wt_i = b_i - a_i$$

ii) Total waiting time, *TWT*: Total waiting time of the jobs,

$$TWT = \Sigma (b_i - a_i)$$

5 Results and Discussion

Fig. 3 demonstrates the performance of HRRN and FCFS algorithms measured by the total waiting time versus number of jobs submitted. The simulation started at job submission rate of 50 jobs per second, and ended with 500 jobs per second with incremental of 50 jobs for each experiment. 10 experiments were conducted in total.

The HRRN achieved smaller total waiting time as compared to FCFS algorithm. From 50 to 150 number of jobs submitted per second, the total waiting time was 0 due to the servers were not fully utilized and the capacity of the jobs submitted were still manageable. At job submission rate of 200 jobs per second, total waiting time was of similarly low for both algorithms due to job sizes were not much in difference. From job submission rate of 250 and above, HRRN outperformed FCFS algorithm all the way till termination of service.

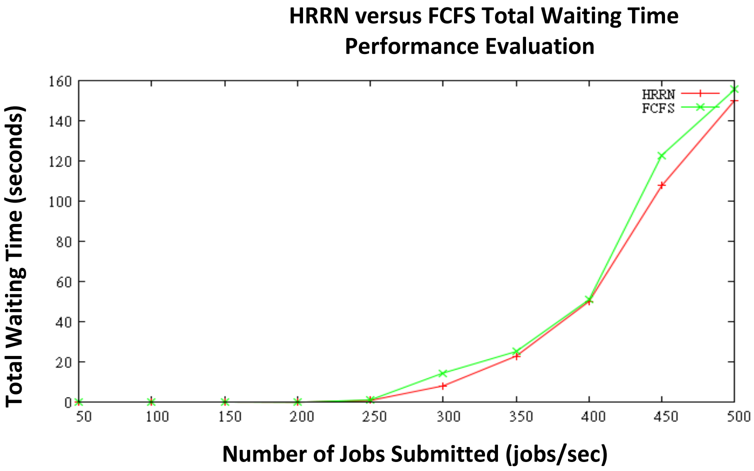


Fig. 3. Simulation result of HRRN and FCFS

The percentage of out performances at 250, 300, 350, 400, 450 and 500 jobs submitted per second were 25%, 6%, 2%, 1%, 15% and 5% respectively. This was mainly due to variety job sizes submitted and heterogeneous servers processing the jobs.

6 Conclusion and Future Works

This paper primarily focuses on scheduling the jobs in a fair and optimized manner. The scheduling algorithms of HRRN and FCFS implemented in this paper reorder jobs in the global queue based on the priority scheme variable. Based on the findings, it can be concluded that HRRN outperform FCFS scheduling algorithm by 5% on average in regards to total waiting time versus number of jobs submitted due to its fair aging priority scheme policy.

Future work should look into implementing HRRN scheduling algorithm in both global and local queues by taking into consideration the average waiting time rather than total waiting which produces relatively small improvement.

References

1. Dafouli, E., Kokkinos, P., Emmanouel, A.V.: Distributed and parallel systems, 93–104 (2008)
2. Martincová, P., Zábovský, M.: Comparison of simulated Grid scheduling algorithms. SYSTÉMOVÁ INTEGRACE 4, 69–75 (2007)
3. Ma, Y.: Design, Test and Implement a Reflective Scheduler with Task Partitioning Support of a Grid, Ph. D.Thesis, Cranfield University (2008)
4. Dong, F.P., Akl, S.G.: Scheduling Algorithms for Grid Computing: State of the Art and Open Problems. Technical Report, School of Computing, Queen’s University: Ontario (2006)
5. Pinedo, M.L.: Planning and Scheduling in Manufacturing and Services, Math. Meth. Oper. Res. 63, 187–189 (2006)
6. Dorigo, M., Di Caro, G.: The Ant Colony Optimization Meta-Heuristic. In: Corne, D., Dorigo, M., Glover, F. (eds.) New Ideas in Optimization, McGraw-Hill, New York (1999)
7. Abraham, S., Baer Galvin, P., Gagne, G.: Operating System Concepts, 7th edn. John Wiley & Sons, Chichester (2005)
8. Elwalid, A., Mitra, D., Widjaja, I.: Routing and protection in GMPLS networks: From shortest paths to optimized designs. IEEE Journal on Lightwave Technology 21(11), 2828–2838 (2003)
9. Kretsis, A., Kokkinos, P., Varvarigos, E.A.: Developing scheduling policies in gLite Middleware. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, Tsukuba (2008)
10. Leemis, H.L., Park, S.K.: Discrete-Event Simulation: A First Course. Prentice Hall, Englewood Cliffs (2004)

Reduced Support Vector Machine Based on k-Mode Clustering for Classification Large Categorical Dataset

Santi Wulan Purnami[1,2], Jasni Mohamad Zain[1], and Abdullah Embong[1]

[1] Faculty of Computer System and Software Engineering, University Malaysia Pahang,
Lebuh Raya Tun Abdul Razak 26300, Kuantan Pahang, Malaysia
[2] Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Keputih, Sukolilo, Surabaya 60111, Indonesia
santipurnami@yahoo.com, jasni@ump.edu.my, ae@ump.edu.my

Abstract. The smooth support vector machine (SSVM) is one of the promising algorithms for classification problems. However, it is restricted to work well on a small to moderate dataset. There exist computational difficulties when we use SSVM with non linear kernel to deal with large dataset. Based on SSVM, the reduced support vector machine (RSVM) was proposed to solve these difficulties using a randomly selected subset of data to obtain a nonlinear separating surface. In this paper, we propose an alternative algorithm, *k*-mode RSVM (KMO-RSVM) that combines RSVM with *k*-mode clustering technique to handle classification problems on categorical large dataset. In our experiments, we tested the effectiveness of KMO-RSVM on four public available dataset. It turns out that KMO-RSVM can improve speed of running time significantly than SSVM and still obtained a high accuracy. Comparison with RSVM indicates that KMO-RSVM is faster, gets smaller reduced set and comparable testing accuracy than RSVM.

Keywords: *k*-mode clustering, smooth support vector machine, reduced support vector machine, large categorical dataset.

1 Introduction

Many Support Vector Machine (SVM) algorithms have been developed to improve testing accuracy of classifiers. Among them, one of the effective algorithms is Smooth Support Vector Machine (SSVM) proposed by Lee, Y.J. and O.L. Mangasarian (2001) [13]. The SSVM algorithm generates a SVM classifier by solving a strongly convex unconstrained minimization problem without using any optimization package. This method uses a Newton Armijo algorithm that can be shown to converge globally and quadratically to the unique solution of the SSVM [12].

The SSVM methods only work well on small to moderate dataset. In practice for large data, for example for solving classification problem (face classification) with number of samples 6977, SSVM cannot be used [2]. Therefore, many of the techniques developed to reduce the amounts of data. Variants methods of reduced set selection for Reduced Support Vector Machines (RSVM) were proposed, such as

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 694–702, 2011.

RSVM based on random sampling [14], RSVM based on stratified sampling [17], and Systematic Sampling RSVM (SSRSVM) that selects the informative data points to form the reduced set [2, 3] and Clustering RSVM (CRSVM) that builds the model of RSVM via Gaussian kernel construction [2, 16].

In this paper, we propose an alternative algorithm called K-Mode RSVM (KMO-RSVM) for handling large categorical dataset. The KMO-RSVM algorithm reduces support vectors by combining *k*-mode clustering technique and RSVM.

We briefly outline the contents of this paper. Section 2 gives the main ideas and formulation for RSVM. In section 3, we describe the proposed algorithm. Firstly the outline of *k*-mode clustering is introduced, and then we provide RSVM based on *k*-mode clustering technique to handle large classification problems with categorical dataset. The numerical experiment and results are presented in section 4. Finally, discussion and conclusions are given in section 5.

2 Reduced Support Vector Machine (RSVM)

In this section, the fundamental concept and main idea of RSVM are described briefly.

2.1 Fundamental Concept

A word about notation and background material is given here. All vector x are column vectors and x' denotes the transpose of x. For a vector $x \in R^n$, the plus function x_+ is defined as $(x_+) = \max\{0, x_i\}$, $i = 1, \dots, n$. For a matrix $A \in R^{mxn}, A_i$ is the i^{th} row of A which is a row vector in R^n. A column vector of one of arbitrary dimensions will be denoted by e. The base of the natural logarithm will be denoted by ε. The p-norm of x will be denoted by $\|x\|_p = \left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}$. For $A \in R^{mxn}$ and $B \in R^{nxl}$, the kernel $K(A, B)$maps $R^{mxn}XR^{nxl}$ into R^{mxl}.

2.2 Review of RSVM

Support vector machines (SVM) have been applied in many classification problems successfully. However, it is restricted to work well on a small dataset. In large scale problems, the full kernel matrix will be very large so it may not be appropriate to use the full kernel matrix. In order to avoid facing such a big full kernel matrix, Lee and Mangasarian, 2001 [14] proposed Reduced Support Vector Machines (RSVM). The key idea of RSVM is randomly selecting a portion of data as to generate a thin rectangular kernel matrix. Then it uses this much smaller rectangular kernel matrix to replace the full kernel matrix [15].

Assume that we are given a dataset consisting of m points in n dimensional real space R^n. These data points are represented by $m\ x\ n$ matrix, where the i^{th} row of matrix A, A_i corresponds to the i^{th} data point. Each point in dataset comes with a class label, +1 or -1. Two classes data A_+ and A_- belong to positive (+1) and negative (-1), respectively. A diagonal matrix D_{mxm} with 1 or -1 along its diagonal can be used to specify the membership of each point. In other words, $D_{ii} = \pm$ depending on whether the label of i^{th} data point is +1 or -1. This is a classification problem which aims to find a classifier that can predict the label of new unseen data points.

We now briefly describe the RSVM formulation for binary classification, which is derived from generalized support vector machine (GSVM) [18] and the smooth support vector machines (SSVM) [13]. The SSVM formulation for non linear case is given as follows [13]:

$$\begin{aligned} &\min_{(u,\gamma,y)} \frac{v}{2} y'y + \frac{1}{2}\left(u'u + \gamma^2\right) \\ &\text{s.t. } D\left(K(A,A')Du - e\gamma\right) + y \geq e \\ &\qquad\qquad\qquad\qquad y \geq 0 \end{aligned} \tag{1}$$

Where v is positive number for balancing training error and the regularization term in objective function. The constraint in equation (1), can be written by:

$$y = (e - D(K(A,A')Du - e\gamma)_+ \tag{2}$$

Thus, we can replace y in (1) by (2) and convert the SVM problem (1) into an equivalent SVM which is an unconstrained optimization problem as follows:

$$\min_{u,\gamma} \frac{v}{2} \left\| p\left(e - D\left(K\left(A,A'\right)Du - e\gamma\right), \alpha\right)\right\|_2^2 + \frac{1}{2}\left(u'u + \gamma^2\right) \tag{3}$$

Where the function p is very accurate smooth approximation to x_+ and is defined:

$$p(x,\alpha) = x + \frac{1}{\alpha}\log\left(1 + \varepsilon^{-\alpha x}\right), \alpha > 0 \tag{4}$$

We replace the full kernel matrix $K(A,A')$ with a reduced kernel matrix $K(A,\overline{A}')$, where $\overline{A}$ consists of $\overline{m}$ random column from A, and the problem becomes:

$$\min_{(\overline{u},\gamma)\in R^{\overline{m}+1}} \frac{v}{2} \left\| p\left(e - D\left(K\left(A,\overline{A}'\right)\overline{D}\overline{u} - e\gamma\right), \alpha\right)\right\|_2^2 + \frac{1}{2}\left(\overline{u}'\overline{u} + \gamma^2\right) \tag{5}$$

The reduced kernel method constructs a compressed model and cuts down the computational cost from $O(n^3)$ to $O(\overline{n}^3)$. It has been shown that the solution of reduced kernel matrix approximates the solution of full kernel matrix well [17].

Algorithm 1. RSVM Algorithm

i. Choose a random subset matrix $\overline{A} \in R^{\overline{m}xn}$ of the original data matrix $A \in R^{mxn}$.

ii. Solve modified version of SSVM (9) where A' only is replaced by $\overline{A}'$ with corresponding $\overline{D} \subset D$

$$\min_{(\overline{u},\gamma)\in R^{\overline{m}+1}} \frac{v}{2} \left\| p\left(e - D\left(K\left(A,\overline{A}'\right)\overline{D}\overline{u} - e\gamma\right), \alpha\right)\right\|_2^2 + \frac{1}{2}\left(\overline{u}'\overline{u} + \gamma^2\right)$$

iii. The separating surface is given by (8) with A' replaced by $\overline{A}'$ as follows:

$$K\left(x', \overline{A}'\right)\overline{D}\overline{u} = \gamma$$

iv. A new input is classified into class +1 or -1 depending on whether the step function: $(K(x', \bar{A}')\bar{D}\bar{u} - \gamma)_*$
is +1 or zero respectively.

3 RSVM Based on *k*-Mode Clustering

In this section, the proposed reduced support vector machine for classification large categorical attributes data set based on *k*-modes clustering will be explained. We called KMO-RSVM. Firstly, we describe *k*-modes clustering for categorical data. Then, we describe algorithm of RSVM based on *k*-modes clustering.

3.1 *K*-Mode Clustering

Most previous clustering algorithms focussed on numerical dataset. However, much of the data existed in the databases is categorical, where attribute values cannot be naturally ordered as numerical values.

Various clustering algorithms have been reported to cluster categorical data. He, Z. *et al* (2005) [5] proposed a cluster ensemble for clustering categorical data. Ralambondrainy (1995) [20] presented an approach by using *k*-means algorithm to cluster categorical data. The approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and treat the binary attributes as numeric in the *k*-means algorithm. Gowda and Diday (1991) [4] used other dissimilarity measures based on “position”, “span” and “content” to process data with categorical attributes. Huang (1998) [7] proposed *k*-modes clustering which extend the *k*-means algorithm to cluster categorical data by using a simple matching dissimilarity measure for categorical objects. Recently, Chaturvedi, *et al* (2001) [1] also presented *k*-modes which used a nonparametric approach to derive clusters from categorical data using a new clustering procedure. Huang, Z (2003) [8] has demonstrated the equivalence of the two independently developed *k*-modes algorithm given in two papers (Huang, Z. 1998; Chaturvedi, *et al* 2001). Then, San, O.M., et al (2004) [21] proposed an alternative extension of the *k*-means algorithm for clustering categorical data which called *k*-representative clustering.

In this study, we concern to adopt *k*-Modes clustering algorithm which was proposed by Huang (1998) [7]. This method is based on *k*-means clustering but removes the numeric data limitation. The modification of *k*-means algorithm to *k*-modes algorithm as follows (Huang, Z. 1998) [7]:

i. Using a simple matching dissimilarity measure for categorical objects
ii. Replacing means of clusters by modes
iii. Using a frequency based method to update the modes.

The simple matching dissimilarity measure can be defined as following. Let X and Y are two categorical objects described by *m* categorical attributes. The dissimilarity measure between X and Y can be defined by the total mismatches of the corresponding attribute categories of the two objects. The smaller the number of mismatches is, the more similar the two objects. Mathematically, it can be presented as follows [7]:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{6}$$

Where

$$\delta(x_j, y_j) = \begin{cases} o & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{7}$$

When (6) is used as the dissimilarity measure for categorical objects, the cost function becomes:

$$P(W,Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{i,l} \delta(x_{i,j}, q_{l,j}) \tag{8}$$

Where $w_{i,l} \in W$ and $Q_l = [q_{l,1}, q_{l,2}, \dots, q_{l,m}] \in \boldsymbol{Q}$.

The *k*-modes algorithm minimizes the cost function defined in equation (8). The *k*-modes algorithm consists of the following steps [7]:

1. Select *k* initial modes, one for each cluster.
2. Allocate an object to the cluster whose mode is the nearest to it according to (6).
3. After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.
4. Repeat 3 until no object has changed clusters after a full cycle test of the whole data set.

3.2 KMO-RSVM

K-means is one of the simplest and famous unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori [7]. The disadvantage of this algorithm is that it can only deal with data sets with numerical attribute. Hence, in this work we have proposed an alternative algorithm, we called KMO-RSVM, to build classifier by combining the *k*-Modes clustering technique and RSVM for classification large categorical data.

The RSVM based on *k*-Modes clustering (KMO-RSVM) algorithm is described below.

Algorithm 2. KMO-RSVM Algorithm. Let k_1 be the number of cluster centroids for positive class and k_2 be the number of cluster centroids for negative class.

Step 1. For each class, runs *k*-Modes clustering to find the cluster centroids c^h. Use clustering results to form the reduced set $\bar{A} = [c^1, c^2, \dots, c^m]'$.
Where $m = k_1 + k_2$.

We use the simplest method, *rule thumb* [11] to determine number of cluster centroids k_1 and k_2. The equation as follow:

$$k \approx \left(\frac{n}{2}\right)^{\frac{1}{2}} \tag{9}$$

Step 2. Two parameters (the SSVM regularization parameter v and kernel parameter γ) are selected using Uniform Design method [9].

Step 3. SSVM classifiers are built on reduced data using Newton Armijo algorithm.

Step 4. Get the separating surface is given by:

$$K\left(x', \bar{A}'\right)\bar{D}\bar{u} = \gamma \tag{10}$$

Step 5. A new unseen data point $x \in R^n$ is classified as class +1 if

$$(K(x', \bar{A}')\bar{D}\bar{u} - \gamma) \geq 0$$

Otherwise *x* is classified as class -1.

4 Experiments and Results

To evaluate the effectiveness of KMO-RSVM, we conducted experiments on four categorical attributes dataset. There are breast cancer dataset, tic-tac-toe dataset, chess dataset and mushroom dataset (Table 1). All of datasets were taken from UCI machine repository [19]. We choose these dataset to test this algorithm because all attributes of the data can be treated as categorical.

Table 1. Descriptions of four categorical dataset

Dataset	Number of data	Number of positive class	Number of negative class	Number of attributes
Breast cancer	683	444	239	9
Tic-tac-toe	958	626	332	9
Chess	3196	1669	1527	36
Mushroom	8124	4208	3916	20

We present two performance evaluations which used to evaluate the KMO-RSVM method. There are 10 fold testing accuracy and response time. In order to give a more objective comparison, we run 10 times on each dataset. All our experiments were performed on a personal computer, which utilizes a 2.00 GHz T6600 Intel(R) Core(TM) 2 duo CPU processor and 2 gigabytes of RAM. This computer runs on windows 7 operating system, with MATLAB 7 installed. The results of experiment on four dataset above using KMO-RSVM and SSVM can be presented as following.

Table 2. Testing accuracy and response time of KMO-RSVM and SSVM method on four categorical dataset

Average **(best in bold)** Standard deviation				
Dataset	SSVM		KMO-RSVM	
	Testing Accuracy (%)	Response time *(seconds)*	Testing Accuracy (%)	Response time *(seconds)*
Breast cancer	0.9724 (0.001061)	299.8901 (20.1303)	**0.9725** (0.001049)	**2.1434** (0.05313)
Tic-tac-toe	**0.9979** (0.001232)	845.0536 (44.37456)	0.9833 (0.000483)	**2.7643** (0.107759)
Chess	**0.9948** (0.000374)	13194 (2509.151)	0.97642 (0.001004)	**18.9042** (1.83739)
Mushroom	N/A	N/A	**0.9992** (0.000146)	**90.4494** (2.93452)

In numerical tests (Table 2), we can see that the response time of KMO-RSVM is always really less than the response time of SSVM. This indicates that the KMO-RSVM can improve speed the response time significantly than SSVM method. Even, on mushroom dataset, while the computer ran out of memory to generate the full non linear kernel (SSVM), the KMO-RSVM can solve the problem with the response time 90.4494 seconds. Meanwhile, it can be seen the testing accuracy of KMO-RSVM slightly decreases than SSVM on tic-tac-toe and chess dataset, whereas on breast cancer dataset the testing accuracy of KMO-RSVM slightly increase than SSVM. Table 2 also shows that the standard deviation of testing accuracy and response time of KMO-RSVM is lower than SSVM. It means the KMO-RSVM is more consistent than SSVM algorithm.

In the next work, we compare our results of KMO-RSVM with RSVM.

Table 3. Comparisons accuracy, time and reduced set ($\bar{m}$) of variants methods (SSVM, RSVM, KMO-RSVM)

Dataset	Measure	Method		
		SSVM	RSVM	KMO-RSVM
Breast cancer	Accuracy	97.16	**97.295**	97.250
	Response time	286.9638	2.5818	**2.1434**
	$\bar{m}$	683	35	**26**
Tic-tac-toe	Accuracy	**99.79**	98.392	98.350
	Response time	816.5248	3.1668	**2.76434**
	$\bar{m}$	958	38	**31**
Chess	Accuracy	**99.430**	97.746	97.642
	Response time	12214	21.12251	**18.90418**
	$\bar{m}$	3196	64	**57**
Mushroom	Accuracy	N/A	99.913	**99.919**
	Response time	N/A	**79.99732**	90.44939
	$\bar{m}$	8124	81	**80**

From table 3, generally it can be concluded that:

- The number reduced set of KMO-RSVM smaller than RSVM method.
- The response time of KMO-RSVN less than RSVM method, except on mushroom dataset.
- Obtained testing accuracy of KMO-RSVM and RSVM are comparable.

5 Conclusions and Discussion

In this study, we have proposed an alternative algorithm, KMO-RSVM which combine *k*-mode clustering and RSVM for classification large categorical dataset. The motivation for KMO-RSVM comes from computational difficulty of SSVM in solving large data set. The numerical results show that KMO-RSVM can improve speed response time significantly and can handle classification for large categorical dataset, when the SSVM method ran out of memory (in case: mushroom dataset). We also compare the classification performance of KMO-RSVM with RSVM. Experiments have indicated that the response time of KMO-RSVN was less than RSVM method, except on mushroom dataset. Testing accuracy of KMO-RSVM and RSVM was found to be comparable.

From the results above it can be concluded that the KMO-RSVM appears to be a very promising method for handling large classification problems especially for categorical dataset.

References

1. Chaturvedi, A., Green, P., Carrol, J.: K-Modes Clustering. Journal of Classification 18, 35–55 (2001)
2. Chien, L.J., Chang, C.C., Lee, Y.J.: Variant Methods of Reduced Set Selection for Reduced Support Vector achines. Journal of Information Science and Engineering 26(1) (2010)
3. Chang, C.-C., Lee, Y.-J.: Generating the reduced set by systematic sampling. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 720–725. Springer, Heidelberg (2004)
4. Gowda, K.C., Diday, E.: Symbolic clustering using a new dissimilarity measure. Pattern Recognition Letters 24(6), 567–578 (1991)
5. He, Z., Xu, X., Deng, S.: A cluster ensemble for clustering categorical data. Information Fusion 6, 143–151 (2005)
6. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of The First Pacific sia Knowledge Discovery and Data Mining Conference, Singapore (1997)
7. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining nd Knowledge Discovery 2, 283–304 (1998)
8. Huang, Z.: A Note on K-modes Clustering. Journal of Classification 20, 257–261 (2003)
9. Huang, C.M., Lee, Y.J., Lin, D.K.J., Huang, S.Y.: Model Selection for Support Vector Machines via Uniform esign. A Special issue on Machine Learning and Robust Data Mining of Computational Statistics and Data Analysis 52, 335–346 (2007)

10. Hsu, C.W., Chang, C.C., Lin, C.J.: Practical Guide To Support Vector Classification. Department of Computer cience and Information Engineering National Taiwan University (2003), http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
11. Mardia, K., et al.: Multivariate Analysis. Academic Press, London (1979)
12. Lee, Y.J.: Support vector machines in data mining. PhD thesis. University of Wisconsin-Madison, USA (2001)
13. Lee, Y.J., Mangasarian, O.L.: A Smooth Support Vector Machine. J. Comput. Optimiz. Appli. 20, 5–22 (2001)
14. Lee, Y.J., Mangasarian, O.L.: RSVM: Reduced Support Vector Machines. In: Proceedings of the First SIAM International Conference on Data Mining. SIAM, Philadelphia (2001)
15. Lee, Y.J., Huang, S.Y.: Reduced Support Vector Machines: A Statistical Theory. IEEE Trans.Neural Network. 18(1) (2007)
16. Jen, L.-R., Lee, Y.-J.: Clustering model selection for reduced support vector machines. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 714–719. Springer, Heidelberg (2004)
17. Lin, K.M., Lin, C.J.: A Study on Reduced Support Vector Machines. IEEE Trans.Neural Network. 14(6), 1449–1459 (2003)
18. Mangasarian, O.L.: Generalized Support Vector Machines. In: Smola, A., Bartlett, P., Scholkopf, B., Schurrmans, D. (eds.) Advances in large Margin Classifiers, pp. 35–146. MIT Press, Cambridge (2000); ISBN: 0-262-19448-1
19. Newman, D.J., Hettich, S., Blake, C.L.S., Merz, C.J.: UCI repository of machine learning database. Dept. of Information and Computer Science. University of California, Irvine (1998), http://www.ics.uci.edu/~mlearn/~MLRepository.html
20. Ralambondrainy, H.: A conceptual version of the K-Means algorithm. Pattern Recognition Letters 16, 1147–1157 (1995)
21. San, O.M., Huynh, V.N., Nakamori, Y.: An alternative extension of the k-means algorithm for clustering categorical data. International Journal Applied Mathematic Computing Science 14(2), 241–247 (2004)
22. Vapnik, V.: The Nature of Statistical Learning Theory, 2nd edn. Springer, New York (1995)

Fractal Analysis of Surface Electromyography (EMG) Signal for Identify Hand Movements Using Critical Exponent Analysis

Angkoon Phinyomark[1], Montri Phothisonothai[2], Pornpana Suklaead[1], Pornchai Phukpattaranont[1], and Chusak Limsakul[1]

[1] Department of Electrical Engineering, Prince of Songkla University
15 Kanjanavanich Road, Kho Hong, Hat Yai, Songkhla 90112, Thailand
[2] College of Research Methodology and Cognitive Science, Burapha University
Chonburi 20131, Thailand
Angkoon.P@hotmail.com,Montrip@buu.ac.th,Suklaead@hotmail.com,
{Pornchai.P,Chusak.L}@psu.ac.th

Abstract. Recent advances in non-linear analysis have led to understand the complexity and self-similarity of surface electromyography (sEMG) signal. This research paper examines usage of critical exponent analysis method (CEM), a fractal dimension (FD) estimator, to study properties of the sEMG signal and to use these properties to identify various kinds of hand movements for prosthesis control and human-machine interface. The sEMG signals were recorded from ten healthy subjects with seven hand movements and eight muscle positions. Mean values and coefficient of variations of the FDs for all the experiments show that there are larger variations between hand movement types but there is small variation within hand movement. It also shows that the FD related to the self-affine property for the sEMG signal extracted from different hand activities 1.944~2.667. These results have also been evaluated and displayed as a box plot and analysis-of-variance (*p* value). It demonstrates that the FD value is suitable for using as an EMG feature extraction to characterize the sEMG signals compared to the commonly and popular sEMG feature, i.e., root mean square (RMS). The results also indicate that the *p* values of the FDs for six muscle positions was less than 0.0001 while that of the RMS, a candidate feature, ranged between 0.0003-0.1195. The FD that is computed by the CEM can be applied to be used as a feature for different kinds of sEMG application.

Keywords: Electromyography (EMG) signal, fractal dimension (FD), critical exponent analysis method (CEM), root mean square (RMS), feature extraction, human-machine interfaces (HMI).

1 Introduction

The surface electromyography (sEMG) signal is one of the most important electrophysiological signals that are commonly used to measure the activity of the muscle and offer the useful information for study of numerous engineering and medical

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 703–713, 2011.

applications [1]. In the engineering applications, use of the sEMG signal as an effective control signal for the control of the prosthesis and other assistive devices are currently increase interested development [2]. This is due to the fact that the usage of the sEMG signal is very easy, fast and convenient. In order to use the sEMG signal as a control signal, feature extraction method should be done before classification step, because raw sEMG signals contain a lot of hidden information in a huge data.

Feature extraction is a method to preserve the useful sEMG information and discard the unwanted sEMG part [3]. Over the last fifteen years, lots of methods have been used and we can classify these methods into four main groups, i.e., time domain, frequency domain, time-frequency or time-scale, and non-linear [2-4]. Time domain and frequency domain features are the first and the second groups that are used to describe characteristic of the sEMG signals based on an analysis of the mathematical functions with respect to time information and frequency information, respectively. The well-known time domain features are integrated EMG (IEMG), root mean square (RMS) and zero crossing (ZC), and the popular frequency domain features are mean frequency (MNF) and median frequency (MDF) [3]. The IEMG, RMS, and ZC features have been widely used in control of the assistive and rehabilitative devices, while the MNF and MDF features are normally used in muscle fatigue analysis; moreover, it has a poor performance for using as the control signal [4]. However, time domain features have been successful to some limits because these methods assume that the sEMG signal is stationary, while the sEMG signal is non-stationary. The third group, time-scale features including, short time Fourier transform (STFT), discrete wavelet transform (DWT) and wavelet packet transform (WPT), have been proposed. The good ability in class separation of features in this group has been presented in numerous research papers [3-5]. However, the drawbacks of their implementation complexity and processing time are main limitation of features in this group [6]. In addition, features in the first three groups extracted features based on linear or statistical analysis but the sEMG signal properties are a complex, non-linear, non-periodic, and non-stationary [1, 7-8]. Therefore, an advancement of the fourth feature group that extracts feature based on non-linear analysis has been interested.

Non-linear feature is essential to extract complexity and self-similarity information of electro-physiological signals [7-9]. Many non-linear analysis methods were found in analysis of the sEMG signal such as approximate entropy, sample entropy, Lempel-Ziv complexity, Lyapunov exponent, Hurst exponent and fractal dimension (FD) [7-12]. However, the FD parameter has been achieved in many research papers. There are many FD estimators such as Higuchi's method, Katz's method, box-counting method and correlation dimension [10-13]. In this research paper, we investigated an advanced FD estimator, namely critical exponent analysis method (CEM). This method is investigated by Nakagawa [14] as an extraction tool for identifying the physiological data. Generally, the CEM has been established as an important tool for detecting the FD parameter of self-affine time series information. It can determine the FD and Hurst exponent (H), which is calculated respect to frequency [14]. It obtains critical exponent of the moment of power spectral density (PSD). The CEM has been successfully applied to various fields such as distinguish vocal sound, identify speaker, fish swimming behavior and complex biomedical signals, including electroencephalogram (EEG) signal [15-21]. However, the CEM has never been applied for an analysis of the sEMG signal. Therefore, the CEM may be a useful tool to characterize the properties of the sEMG signal. Moreover, the limitations of the FD estimators based on time domain have been reported in a number of research papers [11, 13]. Thus the FD estimators based on frequency domain maybe provided the distinguishable and valuable information.

The aim of this paper presents the usage of the CEM to study the non-linear properties of the sEMG signal and to use these properties to identify various kinds of hand movements. As a result, this parameter can be used as a feature parameter for the recognition system. To demonstrate its performance, the RMS feature is used as a comparative candidate in this study due to a widely use and classification ability of this feature. This paper is organized as follows. In Section 2, the experiments and data acquisition are introduced in details. After that the proposed method, CEM, is defined in this section with the evaluation functions. In Section 3, the results of CEM and RMS features are reported and discussed. Finally, the concluding remarks are drawn in Section 4.

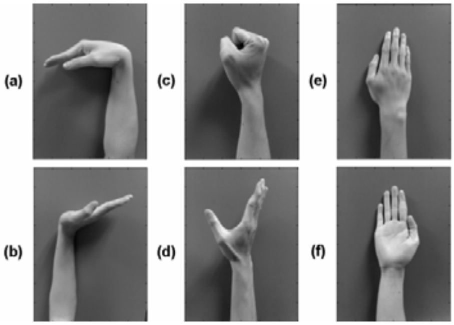


Fig. 1. Six different kinds of hand movements (a) wrist flexion (b) wrist extension (c) hand close (d) hand open (e) forearm pronation (f) forearm supination [22]

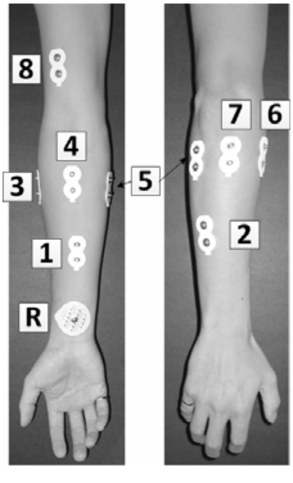


Fig. 2. Eight muscle positions on the right arm of the subject 1 [23]

2 Materials and Methods

2.1 Experiments and Data Acquisition

The experimental setup and protocol of the sEMG signals that were used in this research paper is explained in this section. Ten healthy subjects volunteered to participate in this study. To determine the relationship between the FDs obtained from CEM and muscle activities, each participant was asked to perform six kinds of hand movements including wrist flexion (WF), wrist extension (WE), hand close (HC), hand open (HO), forearm pronation (FP) and forearm supination (FS) with rest state (RS), as shown in Fig. 1. The sEMG signals were recorded from different muscle positions with varied complexity. Seven muscle positions on the right forearm and a muscle position on the upper-arm were selected to perform the isometric contraction activities. All of the eight positions are presented in Fig. 2. The subject was asked to maintain each movement in a time period of three seconds per time. In a trial, each movement was repeated for four times. To measure and evaluate the fluctuation of the sEMG signals, six trials were contained for each session within a day and four sessions on four separate days were acquired for each subject. In the total 5,376 data sets with three-second period were obtained for each subject which is enough to confirm the ability of the proposed method.

The sEMG signals were recorded by Duo-trode Ag/AgCl electrodes (Myotronics, 6140) for each muscle. The wrist was used as the reference electrode position by an Ag/AgCl Red Dot electrode (3M, 2237). The sEMG signals were amplified by the differential amplifiers (Model 15, Grass Telefactor) that were set a gain of 1000 and bandwidth of 1-1000 Hz. The sampling frequency rate was set at 3000 Hz using an analog-to-digital converter board (National Instruments, PCI-6071E). More details of experiments and data acquisition are described in Ref. [23].

2.2 Critical Exponent Analysis Method

Fractal analysis is a mathematical tool for handling with the complex systems. Therefore, a method of the estimating FD has been found a useful for an analysis of biomedical signals. Moreover, the FD can quantity of the information embodied in the signal pattern in terms of morphology, spectrum and variance. This research paper proposed the FD evaluation based on the CEM [14] and used the RMS as a comparative feature.

Denoting the PSD $P_H(v)$ of the observed signals in frequency domain and the frequency of the sEMG signal v. If the PSD satisfies a power law due to self-affinity characteristic of the sEMG signal which can be defined as:

$$P_H(v) \sim v^{2H+1} = v^{-\beta}. \tag{1}$$

In the CEM, the α is the moment exponent and the α^{th} moment of the PSD (I_α) is determined as

$$I_\alpha = \int_1^U P_H(v)v^\alpha dv, \quad (-\infty < \alpha < \infty), \tag{2}$$

where the upper limit of the integration *U* corresponds to the highest considered frequency and the normalized frequency *v* whose lower cut-off corresponds to one. Here *α* is a real number. If we consider the limited frequency bands and substitute Eq. (1) into Eq. (2) thus the equation was given as

$$I_\alpha \sim \int_1^U v^{\alpha-\beta} dv = \frac{1}{\alpha-\beta+1}(U^{\alpha-\beta+1} - 1), \tag{3}$$

$$= \int_1^U v^{X-1} dv = \frac{1}{X}(U^X - 1), \tag{4}$$

$$= \frac{2}{X}\exp\left(\frac{uX}{2}\right)\sinh\left(\frac{uX}{2}\right), \tag{5}$$

where *X* and *u* are defined as

$$X = \alpha - \beta + 1, \tag{6}$$

and

$$u = \log U, \tag{7}$$

respectively. Thus by taking the logarithm of moment I_α and differentiating it to the third order, the formula can be written as

$$\frac{d^3}{d\alpha^3}\log I_\alpha = \frac{I'''_\alpha I_\alpha^2 - 3I''_\alpha I'_\alpha I_\alpha + 2(I'_\alpha)^3}{I_\alpha^3}, \tag{8}$$

$$= \frac{2}{8}u^3 \operatorname{cosech}^3\left(\frac{uX}{2}\right)\cosh\left(\frac{uX}{2}\right) - \frac{2}{X^3}. \tag{9}$$

We then determine the critical value $\alpha = \alpha_c$ at which the value of the third order derivative of log I_α with respect to is zero $d^3\log I_\alpha/d\alpha^3 = 0$. Finally, from this value of α_c, $\beta = \alpha_c-1$. The estimated FD is given by

$$FD = 2 - \frac{\alpha_c}{2} = 2 - H. \tag{10}$$

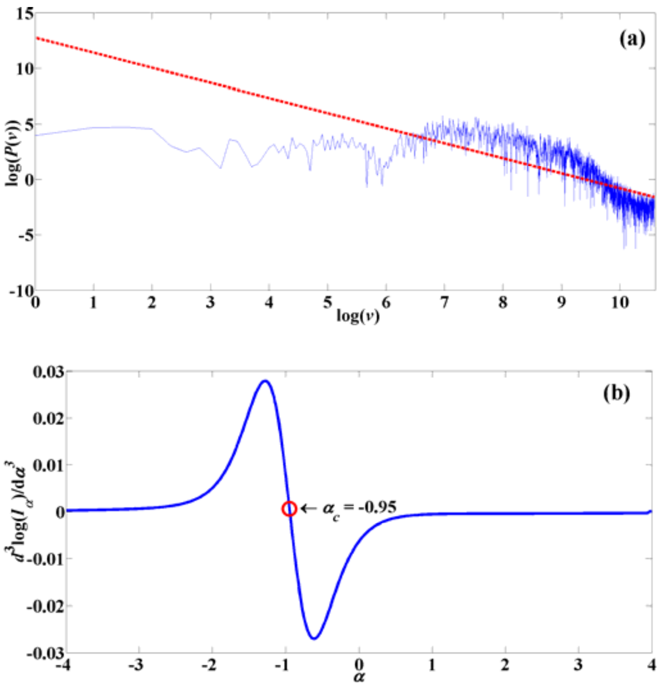


Fig. 3. (a) Log-log plot of PSD $P_H(v)$ versus frequency v - (b) Third order derivative of the logarithmic function and the zero crossing point - of the WF movement and muscle position 1 of the subject 1. Note that in (a), the solid line is $P_H(v)$ and the diagonal dash line is the slope of the log-log plot of the $P_H(v)$ which estimates by linear regression

In this study, to implement the CEM according to Eq. 8, we set the step size of the moment exponent, which had been $\alpha_\Delta = 0.01$. To show the calculating step of the CEM, the sEMG data with the WF movement and the muscle position 1 of the first subject was used and presented in Fig. 3

2.3 Root Mean Square Method

RMS is a time domain feature that is related to the constant force and non-fatiguing contraction. This feature is a widely useful feature in the sEMG pattern classification [2-4, 22-23]. Generally, it similar to standard deviation method and also give the same information for the other commonly used time domain features such as IEMG, variance and mean absolute value, which can be expressed as

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}\,, \tag{11}$$

where x_n represents the sEMG signal in a segment and N denotes the length of the sEMG signal.

In the experiment, the sEMG data during hand movements in each action repeat were processed. To evaluate the sEMG feature respect to time, we set a window function as a rectangular type in both sEMG features. The window size is set to 3,072 points (approx. 1 second) and the moving window with intervals is set to 1024 points (approx. 1/3 second) in this study.

2.4 Evaluation Functions

The selection of sEMG feature extraction is an important stage to achieve the optimal performance in signal pattern classification. In [4] the capability of sEMG feature is qualified by three properties, i.e., class separability, robustness, and complexity. In this research paper, we are demonstrating a usefulness of a novel applied feature, the CEM to identify different kinds of hand movements, especially in class separability point of view. A high capability of class separability is reached when separation between movement classes is high and variation within movement class is also small. To evaluate the first condition, we calculated mean values of the FD of CEM and RMS for each movement and muscle position. We can observe the discrimination between seven types of hand movements and eight kinds of muscle positions in the space of features.

To evaluate the second condition, coefficient of variance (CV) was computed. The CV is a useful method to measure of the dispersion of a probability distribution. It is similar to standard deviation (SD) method, a widely used measurement of variability, and it is defined as the ratio of the SD to the mean μ, which can be expressed as $CV = SD/\mu$. Both CV and SD indices show how much variation or dispersion there is from the mean value. A low CV indicates that feature values tend to be very close to the mean, whereas high CV indicates that the features are spread out over a large range of values. The advantage of the CV index over the SD index is when comparing between feature values with different units or largely different means. Therefore, usage of the CV for comparison instead of the SD is suggested.

These results can be also evaluated by using box plot and analysis-of-variance (ANOVA) method. We can observe the performance of class separability from the box range in the box plot. The ANOVA is a significant statistical test to obtain the p value that is the one way to demonstrate distinguishes between classes. When the p value is smaller than 0.005 or 0.001, it commonly means that significant different between their means are obtained. In this study, we compared the ability of the FD feature with the popular feature, the RMS by using the above introduced evaluation indexes.

3 Results and Discussion

In this study, a large sEMG data has been used and evaluated. Mean value of the FD and the RMS features and CV of their values for each movement and muscle position from all subjects are shown in Table 1 through Table 4, respectively. From the values of the respective FDs that are summarized in Table 1, we found that the FD related to self-affine property for the sEMG signals. Since each subject will present different sEMG patterns during hand movements. The FD values range between 1.944~2.667. In addition, a higher FD indicates a more complicated structure of the sEMG signals or the quantity of information embodied in a pattern.

Table 1. Mean values of the FDs estimated by using CEM at each movement from 10 subjects and 8 muscle positions

Muscle				Movements			
position	WF	WE	HC	HO	FP	FS	RS
1	2.2119	2.3001	2.3215	2.1972	2.1669	2.1816	2.1608
2	2.3810	2.3162	2.3281	2.4636	2.2546	2.2871	2.2257
3	2.2731	2.3182	2.1973	2.4609	2.1975	2.2222	2.1933
4	2.1778	2.1835	2.3045	2.2207	2.1416	2.1944	2.1398
5	2.3153	2.2702	2.3756	2.2240	2.2491	2.1579	2.1563
6	2.4173	2.3092	2.2840	2.4991	2.2907	2.3111	2.2376
7	2.2957	2.3494	2.3039	2.3628	2.2407	2.2490	2.2009
8	2.2538	2.2438	2.2737	2.2261	2.3150	2.1519	2.2415

Table 2. Mean values of RMS at each movement from 10 subjects and 8 muscle positions

Muscle				Movements			
position	WF	WE	HC	HO	FP	FS	RS
1	0.2319	0.2833	0.3679	0.2442	0.3222	0.3373	0.2083
2	0.2280	0.1908	0.2343	0.4373	0.2428	0.2608	0.1347
3	0.1810	0.2129	0.1770	0.4663	0.2201	0.2350	0.1385
4	0.1878	0.1769	0.3524	0.2142	0.2402	0.2793	0.1616
5	0.2790	0.2196	0.3995	0.1981	0.3005	0.2478	0.1584
6	0.3881	0.2825	0.2833	0.7373	0.3924	0.3664	0.2099
7	0.1843	0.2652	0.2317	0.2820	0.2508	0.2419	0.1389
8	0.1719	0.1682	0.2038	0.1862	0.2835	0.1765	0.1508

Table 3. Average values of CVs of the FDs estimated by using CEM at each movement from 10 subjects and 8 muscle positions

Muscle				Movements			
position	WF	WE	HC	HO	FP	FS	RS
1	0.0362	0.0395	0.0375	0.0372	0.0449	0.0426	0.0366
2	0.0387	0.0368	0.0409	0.0396	0.0490	0.0507	0.0404
3	0.0377	0.0411	0.0366	0.0380	0.0410	0.0451	0.0353
4	0.0377	0.0373	0.0387	0.0399	0.0446	0.0430	0.0351
5	0.0365	0.0387	0.0393	0.0394	0.0503	0.0399	0.0389
6	0.0401	0.0418	0.0425	0.0380	0.0477	0.0490	0.0440
7	0.0377	0.0396	0.0402	0.0416	0.0449	0.0468	0.0382
8	0.0462	0.0470	0.0502	0.0495	0.0528	0.0429	0.0435

Table 4. Average values of CVs of RMS at each movement from 10 subjects and 8 muscle positions

Muscle	Movements						
position	WF	WE	HC	HO	FP	FS	RS
1	0.8937	0.6953	0.6942	0.9261	1.0078	0.9105	0.9688
2	0.5643	0.6539	0.6069	0.3704	0.8250	0.6753	0.9362
3	0.7266	0.6106	0.8942	0.4207	0.9460	0.8208	0.9339
4	0.8973	0.8178	0.6354	0.7775	0.9852	0.8831	0.9219
5	0.5525	0.6800	0.5353	0.7626	0.7207	0.9264	0.9229
6	0.5506	0.7433	0.7323	0.3642	0.8384	0.7203	0.9366
7	0.6747	0.5746	0.6362	0.5013	0.8151	0.7418	0.8354
8	0.4825	0.4922	0.5104	0.5299	0.4621	0.7212	0.4762

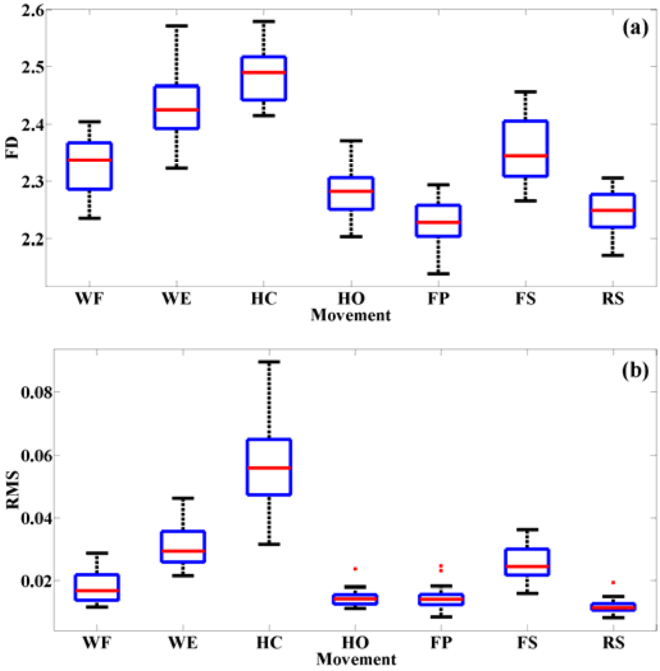


Fig. 4. Box plot of (a) FD (b) RMS from 7 hand movements of the subject 1 and muscle position 1

From the observed comparison between mean values of CEM and RMS in Table 1 and Table 2, it showed that CEM feature contains more separation than RMS feature. Moreover, CV values of CEM feature have the smaller than ten times of CV values of RMS feature, as can be observed in Table 3 and Table 4. To be easily observed, a case of the FD and RMS from seven hand movements of the first muscle position of the subject 1 is presented in Fig. 4 with a box plot. In addition, the *p* values conducted from ANOVA were used to determine significance of the separation of feature to identify different hand movements. From Table 5, it is observed that the FD feature is more significant than that of the RMS feature. The *p* value of the FD for six muscle positions was less than 0.0001 while that of the RMS ranged between 0.0003-0.1195. In additions, it has been found that the FD has direct relation to the electrode position. From the results, the positions of muscle 2 to 7 are suitable for discriminating hand movements. However, the first and the last positions of the muscles are not suitable for both features.

Since each subject will present different sEMG patterns during hand movements. Therefore, the significant of the *p* value shall be observed for each subject. In this study, we show numbers of the subjects who have the average *p* value more than 0.001, in Table 6. It shows that the RMS feature has no significance at least one subject from each muscle position. In the other hand, the FD feature has only four cases from muscle positions 1 and 8 that have no significance at the 0.1% statistically significant level. This clearly demonstrates that the FD is a more reliable feature for classifying hand movements.

Table 5. The average *p* value of 10 subjects for all 7 hand movements and 8 muscle positions

Feature	Muscle positions							
	1	2	3	4	5	6	7	8
FD	0.0035	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0586
RMS	0.0124	0.0241	0.0007	0.0158	0.0094	0.0003	0.0447	0.1195

Table 6. The number of subject that has the average *p* value more than 0.001 from 10 subjects

Feature	Muscle positions							
	1	2	3	4	5	6	7	8
FD	1	0	0	0	0	0	0	3
RMS	3	2	1	4	2	1	2	5

4 Conclusion

Fractal analysis of sEMG signals from different hand movements has been proposed. We evaluated its FD based on CEM with time dependence method. The critical exponent value characterizes the self-affine property of the sEMG signal. The method described in this paper can be considerably utilized the sEMG-based classifications and its performance shows a better than another popular sEMG feature, namely RMS. In future works, use of FD estimated by CEM as an input parameter for recognition system shall be evaluated by different types of classifiers such as artificial neural network and linear discriminant analysis. Moreover, the optimal parameters of the CEM should be evaluated such as window length, increment step of exponent α, etc.

Acknowledgments. This work was supported in part by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0110/2550), and in part by NECTEC-PSU Center of Excellence for Rehabilitation Engineering, Faculty of Engineering, Prince of Songkla University. The authors also gratefully acknowledge the support of Dr. Adrian D.C. Chan from the Carleton University, Canada, for providing the EMG data.

References

1. Merletti, R., Parker, P.: ELECTRO-MYOGRAPHY Physiology, Engineering, and Noninvasive Applications. John Wiley & Sons, Inc, Hoboken (2004)
2. Zecca, M., Micera, S., Carrozza, M.C., Dario, P.: Control of multifunctional prosthetic hands by processing the electromyographic signal. Crit. Rev. Biomed. Eng. 30, 459–485 (2002)
3. Oskoei, M.A., Hu, H.: Myoelectric control systems—A survey. Biomed. Signal Process. Control 2, 275–294 (2007)
4. Boostani, R., Moradi, M.H.: Evaluation of the forearm EMG signal features for the control of a prosthetic hand. Physiol. Meas. 24, 309–319 (2003)
5. Englehart, K., Hudgins, B., Parker, P.A.: A wavelet-based continuous classification scheme for multifunction myoelectric control. IEEE Trans. Biomed. Eng. 48, 302–311 (2001)
6. Tkach, D., Huang, H., Kuiken, T.A.: Study of stability of time-domain features for electromyographic pattern recognition. J. NeuroEng. Rehabil. 7, 21 (2010)
7. Lei, M., Wang, Z., Feng, Z.: Detecting nonlinearity of action surface EMG signal. Phys. Lett. A 290, 297–303 (2001)
8. Meng, Y., Liu, Y., Liu, B.: Test nonlinear determinacy of Electromyogram. In: 27th Annual Conference of the IEEE Engineering in Medicine and Biology, pp. 4592–4595. IEEE Press, New York (2005)
9. Padmanabhan, P., Puthusserypady, S.: Nonlinear analysis of EMG signals – A chaotic approach. In: 26th Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 608–611. IEEE Press, New York (2004)
10. Arjunan, S.P., Kumar, D.K.: Decoding subtle forearm flexions using fractal features of surface electromyogram from single and multiple sensors. J. Neuro. Eng. Rehabil. 7, 53 (2010)
11. Gitter, J.A., Czerniecki, M.J.: Fractal analysis of the electromyographic interference pattern. J. Neurosci. Methods 58, 103–108 (1995)
12. Gupta, V., Suryanarayanan, S., Reddy, N.P.: Fractal analysis of surface EMG signals from the biceps. Int. J. Med. Inf. 45, 185–192 (1997)
13. Hu, X., Wang, Z.-Z., Ren, X.-M.: Classification of surface EMG signal with fractal dimension. J. Zhejiang Univ. Sci. B 6, 844–848 (2005)
14. Nakagawa, M.: A critical exponent method to evaluate fractal dimensions of self-affine data. J. Phys. Soc. Jpn. 62, 4233–4239 (1993)
15. Petry, A., Barone, D.A.C.: Speaker identification using nonlinear dynamical features. Chaos Solitons Fractals 13, 221–231 (2002)
16. De Oliveira, L.P.L., Roque, W.L., Custódio, R.F.: Lung sound analysis with time-dependent fractal dimensions. Chaos Solitons Fractals 10, 1419–1423 (1999)
17. Sabanal, S., Nakagawa, M.: The fractal properties of vocal sounds and their application in the speech recognition model. Chaos Solitons Fractals 7, 1825–1843 (1996)
18. Nimkerdphol, K., Nakagawa, M.: Effect of sodium hypochlorite on Zebrafish swimming behavior estimated by fractal dimension analysis. J. Biosci. Bioeng. 105, 486–492 (2008)
19. Nimkerdphol, K., Nakagawa, M.: 3D locomotion and fractal analysis of Goldfish for acute toxicity bioassay. Int. J. Biol. Life Sci. 2, 180–185 (2006)
20. Phothisonothai, M., Nakagawa, M.: EEG-based fractal analysis of different motor imagery tasks using critical exponent method. Int. J. Biol. Life Sci. 1, 175–180 (2005)
21. Phothisonothai, M., Nakagawa, M.: Fractal-based EEG data analysis of body parts movement imagery tasks. J. Physiol. Sci. 57, 217–226 (2007)
22. Phinyomark, A., Limsakul, C., Phukpattaranont, P.: A novel feature extraction for robust EMG pattern recognition. J. Comput. 1, 71–80 (2009)
23. Chan, A.D.C., Green, G.C.: Myoelectric control development toolbox. In: 30th Conference of Canadian Medical & Biological Engineering Society, M0100 (2007)

Robust Eye Movement Recognition Using EOG Signal for Human-Computer Interface

Siriwadee Aungsakun, Angkoon Phinyomark, Pornchai Phukpattaranont,
and Chusak Limsakul

15 Kanjanavanich Road, Kho Hong, Hat Yai, Songkhla 90112 Thailand
{Siriwadee.A,Angkoon.P}@hotmail.com,
{Pornchai.P,Chusak.L}@psu.ac.th

Abstract. Electrooculography (EOG) signal is one of the useful biomedical signals. Development of EOG signal as a control signal has been paid more increasing interest in the last decade. In this study, we are proposing a robust classification algorithm of eight useful directional movements that it can avoid effect of noises, particularly eye-blink artifact. Threshold analysis is used to detect onset of the eye movements. Afterward, four beneficial time features are proposed that are peak and valley amplitude positions, and upper and lower lengths of two EOG channels. Suitable threshold conditions were defined and evaluated. From experimental results, optimal threshold values were selected for each parameters and classification accuracies approach to 100% for three subjects testing. To avoid the eye-blink artifact, the first derivative was additionally implemented.

Keywords: Electrooculography (EOG) signal, eye movement, eye blink artifact, robust pattern recognition, human-machine interface (HMI).

1 Introduction

Recently, many research works are under way into means of enabling disabled to communicate effectively with machine. Depending on the users' capabilities, different kinds of interfaces have been proposed, such as, vision based head/hand gesture, speech recognition, sip and puff, head or chin controller, ultrasonic non-contact head controller and brain-computer interface [1-6]. However, due to limitations of each interface, for instance, speech recognition and vision based head/hand gesture have a major problem in noisy and outdoor environments or ultrasonic non-contact head controller has a low classification performance, electrooculography (EOG) signal have been proposed to be sufficient signal to be used in human-machine interface (HMI) [7-11]. In this study, we are promoting a usefulness of EOG signal to be used as an efficient hand-free control interface.

EOG signal is commonly used to record activities of human eye. It is a bio-electrical signal that detects changes in eye positions. The EOG signal is generated by the potential difference between the cornea and the ocular fundus. It is ordinarily referred to be called as a cornea-retinal potential (CRP) [12]. This potential difference comes from large presence of the electrically active nerves in the retina equate to the

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 714–723, 2011.

front of the eye. This potential can be considered as a steady electrical dipole with a positive pole at the cornea and a negative pole at the retina [13]. Because of its relatively large amplitude compared to other biomedical signals, which its amplitudes range between 15 and 200 µV. In addition, due to a linear relationship between EOG signal and eye movements and easy detection of waveform, the EOG signal may be look like an ideal candidate for eye movement classification and control system.

EOG signal has been widely and successfully used in biomedical engineering applications, especially in HMIs. Many efficient HMIs have been developed and implemented in the last decade, such as, electrical wheelchair control [7], mobile robot control [8], cursor mouse control [9], eye writing recognition [10], and eye activity recognition [11]. Various techniques have been used to detect and classify the EOG signals and eye movements. Pattern recognition techniques, including neural networks and support vector machine, have been widely used and have been proven their classification performance [14-15]; whereas, their computational times and implemental complexity become a major limitation of these techniques, especially for micro-controller device. In this study, we have been proposed non-pattern recognition algorithm based on time domain feature and threshold analysis to discriminate eight commonly used directional eye movements. Moreover, the proposed technique can be availably implemented for a real-time system and can be used in noisy environment.

2 Materials and Methods

Eight directional eye movements (up, down, right, left, up-right, up-left, down-right, and down-left) are basic movements for most of the HMIs, particularly the first four directions [7-11, 14-15]. Normally, two channel EOG signals have been used to acquire the information from horizontal and vertical eye movements. In this section, we described procedure of the recorded EOG signals and the proposed algorithm to characterize these directional movements.

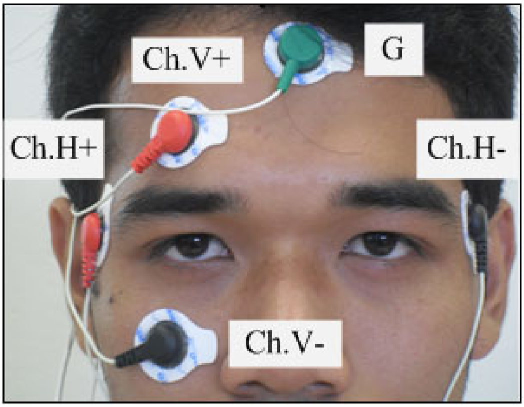


Fig. 1. EOG electrode placements where electrodes Ch.V+ and Ch.V- measure the vertical movements, and Ch.H+ and Ch.H- measure the horizontal movements

2.1 Experiments and Data Acquisition

Independent measurements can be obtained from both eyes. However, in the vertical direction, two eyes move in conjunction; hence, for the vertical signal, only one right eye was used. Five surface electrodes were put around the eyes, as shown in Fig. 1. Vertical leads were acquired on the above and below of the right eye (Ch.V+ and Ch.V-). Horizontal leads were acquired by two electrodes on the right and the left of the outer canthi (Ch.H+ and Ch.H-). A reference electrode was placed on forehead (G). All EOG signal recordings were carried out using a commercial wireless system (Mobi6-6b, TMS International BV, Netherlands). A band-pass filter of 1-500 Hz bandwidth and an amplifier with 19.5x were set for the recorded system. The sampling rate was set to 1024 Hz. However, the energy frequency bands of the EOG signal are fallen in range of 1 to 10 Hz. Thus the sampling rate was reduced to 128 Hz in pre-processing stage.

The EOG data were recorded from three normal subjects with eight directional eye movements: eyes move -down (M1), –up (M2), -left (M3), -right (M4), -down and left (M5), -down and right (M6), -up and left (M7), and -up and right (M8). All of these activities were held two seconds. Each activity was performed five times throughout a trial. As a result, fifteen data sets were obtained from each directional movement.

2.2 Eye Movement Detecting Algorithm

To discriminate eight directional movements mentioned above, we have been proposed a simple and effective non-pattern recognition algorithm based on time domain features and threshold analysis. Our proposed technique can be availably implemented for a real-time system and can be used in noisy environment. The procedure of the proposed algorithm is shown in Fig. 2. In this figure, EOG signal time series ($\{x(i)\}$), position of time samples (i), onset threshold value (THR_{ON}), classified threshold values of upper and lower of vertical V and horizontal H signals (THR_{UV}, THR_{LV}, THR_{UH}, and THR_{LH}), and artifact index (AI) were defined.

Firstly, threshold analysis is used to detect simultaneously starting point of movements from both eyes with suitable level THR_{ON}. Afterward, eight beneficial time features are respectively proposed that consist of peak and valley amplitude positions (PAP_V, VAP_V, PAP_H, VAP_H), and upper and lower lengths (UL_V, LL_V, UL_H, LL_H) of two EOG channels. Definitions of these features are respectively expressed in Figs. 3 and 4. Subsequently, suitable threshold conditions, THR_{UV}, THR_{LV}, THR_{UH}, and THR_{LH}, were defined in Table 1 in order to discriminate eight movements from eight time features which were early calculated. Through experiments, optimal threshold values were evaluated for all parameters. Finally, eight movement classes (M1-M8) were examined for the output parameter (OUT). In addition, when resting and other movements were detected, output OUT is set to M0. To avoid eye-blink artifact, the first derivative of UL_V was additionally implemented and the AI was also calculated with a pre-defined threshold from the first derivative result. When the logical value of AI was defined as true it means that more than one burst signal is found. That is eye-blink artifact was established.

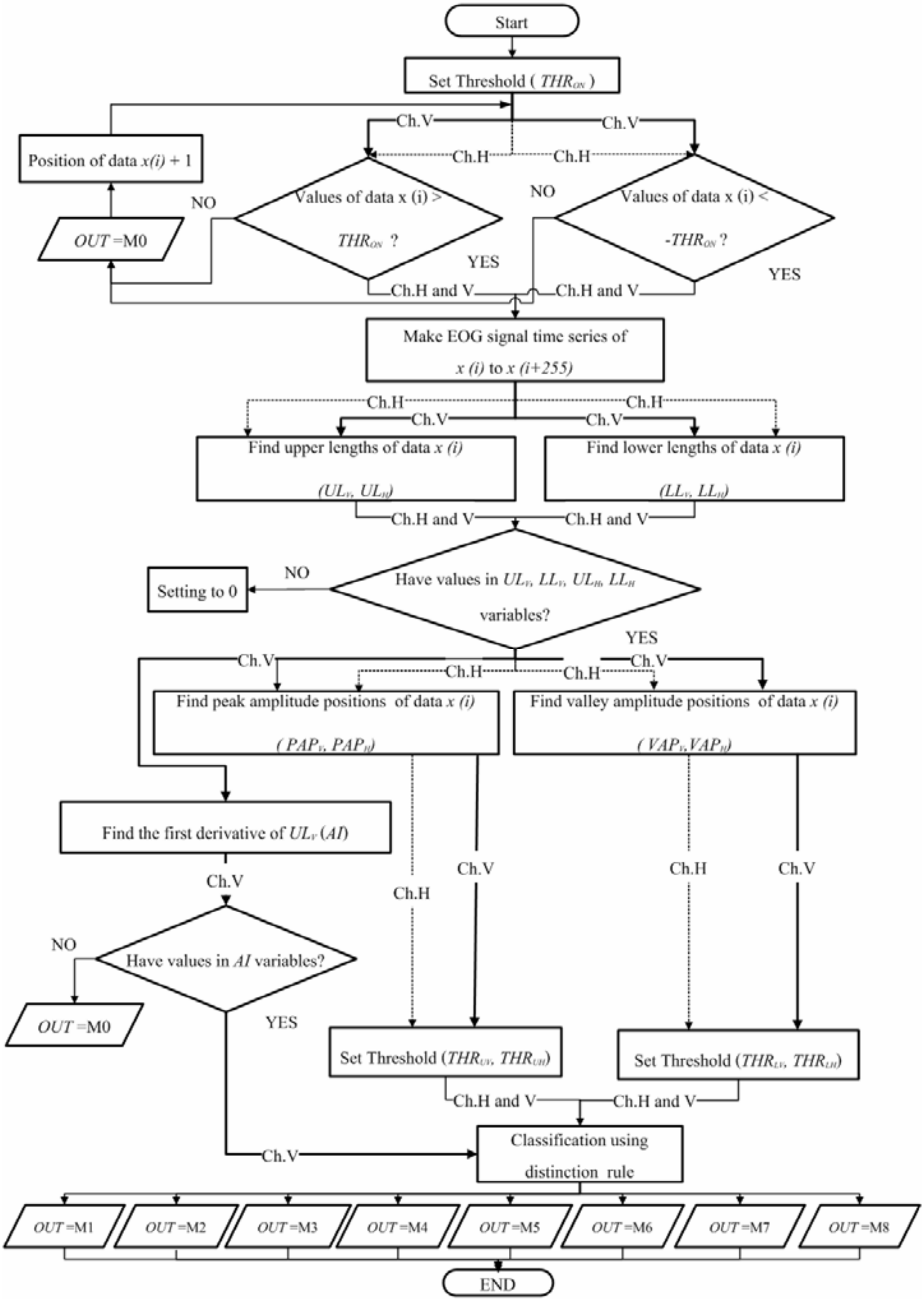

Start
Set Threshold (THR_{ON})
Ch.V
Ch.V
Ch.H
Ch.H
Position of data $x(i)$ + 1
NO
Values of data x (i) > THR_{ON} ?
NO
Values of data x (i) < $-THR_{ON}$?
OUT =M0
YES
YES
Ch.H and V
Ch.H and V
Make EOG signal time series of x (i) to x (i+255)
Ch.H
Ch.H
Ch.V
Ch.V
Find upper lengths of data x (i) (UL_V, UL_H)
Find lower lengths of data x (i) (LL_V, LL_H)
Ch.H and V
Ch.H and V
Setting to 0
NO
Have values in UL_V, LL_V, UL_H, LL_H variables?
YES
Ch.V
Ch.H
Ch.H
Ch.V
Find peak amplitude positions of data x (i) (PAP_V, PAP_H)
Find valley amplitude positions of data x (i) (VAP_V,VAP_H)
Find the first derivative of UL_V (AI)
Ch.V
Ch.H
Ch.V
Ch.H
Ch.V
NO
Have values in AI variables?
OUT =M0
YES
Set Threshold (THR_{UV}, THR_{UH})
Set Threshold (THR_{LV}, THR_{LH})
Ch.H and V
Ch.H and V
Ch.V
Classification using distinction rule
OUT =M1
OUT =M2
OUT =M3
OUT =M4
OUT =M5
OUT =M6
OUT =M7
OUT =M8
END

Fig. 2. Flowchart of the proposed EOG classification algorithm

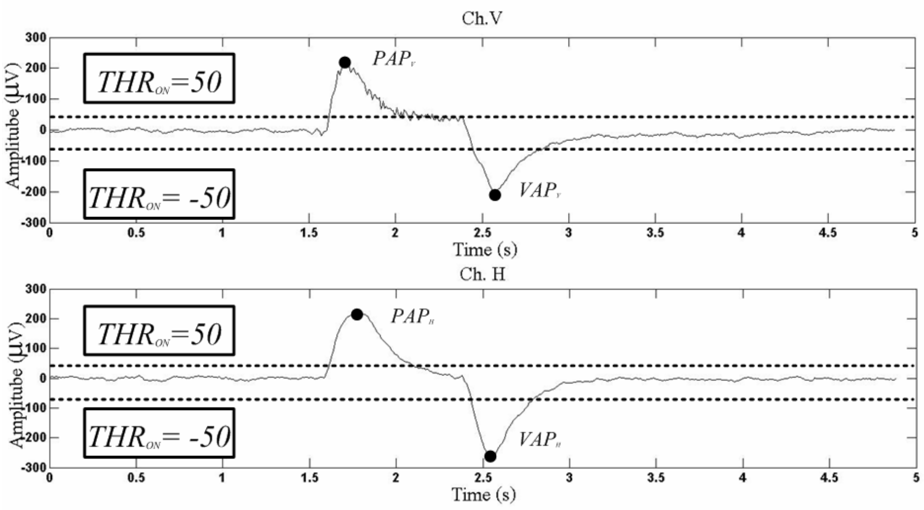


Fig. 3. Parameters of PAP_V, VAP_V, PAP_H and VAP_H.

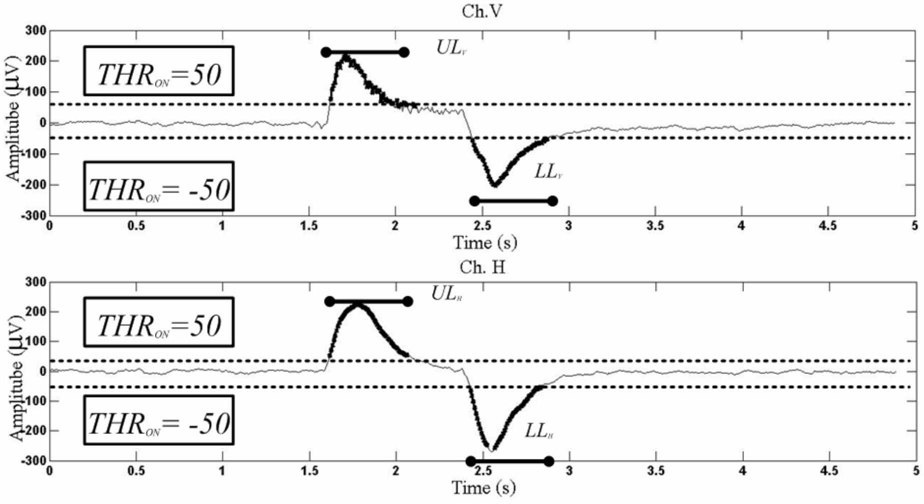


Fig. 4. Parameters of UL_V, LL_V, UL_H, and LL_H**.**

3 Results and Discussion

Firstly, from the observation of amplitude shape, we defined value of the THR_{ON} as 50 μV for detecting simultaneously starting point of movements from both eyes. Afterward, eight time domain features are calculated for two EOG channels. Based on

Table 1. Distinction feature rule

If $PAP_V > VAP_V$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} <= UL_{H,}$ and $THR_{LH} <= LL_{H}$, then OUT = M1	If $PAP_V > VAP_{V,} PAP_H > VAP_H$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} >= UL_{H,}$ and $THR_{LH} >= LL_{H}$, then OUT = M5
If $PAP_V < VAP_V$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} <= UL_{H,}$ and $THR_{LH} <= LL_{H}$, then OUT = M2	If $PAP_V > VAP_{V,} PAP_H < VAP_H$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} >= UL_{H,}$ and$THR_{LH} >= LL_{H}$, then OUT = M6
If $PAP_H > VAP_H$ and $THR_{UV} <= UL_{V,} THR_{LV} <= LL_V$ and $THR_{UH} >= UL_{H,}$ and $THR_{LH} >= LL_{H}$, then OUT = M3	If $PAP_V < VAP_{V,} PAP_H > VAP_H$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} >= UL_{H,}$ and$THR_{LH} >= LL_{H}$, then OUT = M7
If $PAP_H < VAP_H$ and $THR_{UV} <= UL_{V,} THR_{LV} <= LL_V$ and $THR_{UH} >= UL_{H,}$ and $THR_{LH} >= LL_{H}$, then OUT = M4	If $PAP_V < VAP_{V,} PAP_H < VAP_H$ and $THR_{UV} >= UL_{V,} THR_{LV} >= LL_V$ and $THR_{UH} >= UL_{H,}$ and$THR_{LH} >= LL_{H}$, then OUT = M8
	Otherwise OUT = M0

the obtained results in Table 2 through 4, all subjects show that values of our proposed eight features are useful for discriminating eight directional movements. However, the suitable threshold values of THR_{UV}, THR_{LV}, THR_{UH}, and THR_{LH} were dependent on each subject. Finally, eight movement classes (M1-M8) were examined for the output parameter (OUT).

Our proposed algorithm for a robust classification of eight directional movements based on EOG signals has two advantages compared to other publication algorithms. Firstly, the algorithm does not affected by various noises, i.e., involuntary single blink (ISB), involuntary double blinks (IDB), and short-duration eye closed (SDC). Fig. 5. shows the detection of left eye movement (M3) on the top panel using the EOG signals from the vertical and horizontal channels on the middle and bottom panels, respectively. Although, there are ISB, IDB, and SDC noises shown in thick lines generated in vertical leads, our proposed algorithm still can detect the motion with 100% accuracy as shown in the top panel.

Secondly, the proposed algorithm provides high accuracy for the classification of eight directional movements based on EOG signals compared to other publications. Fig. 6. shows the detection of eight eye movements (M1-M8) on the top panel

Table 2. Mean values and standard deviation values of 8 features for subject 1

	Down		Up		Left		Right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_{V}	50.8	6.4	71.4	11.3	12.0	5.1	7.6	12.6
LTL_{V}	56.0	1.7	56.4	14.1	54.2	112.3	4.8	10.7
UTL_{H}	3.2	5.2	3.6	3.4	57.6	8	51.4	1.9
LTL_{H}	-	-	2.4	3.4	62.6	9.5	53	6.2
PAP_{V}	100.8	23.9	13	4.1	144.8	67.7	65.4	96.1
VAP_{V}	14.0	1.0	110	11.1	101.4	133.1	48	107.3
PAP_{H}	3.2	4.4	5.2	5.8	101.8	15.1	12.6	0.5
VAP_{H}	-	-	-	-	14	1.6	102.8	12.6
	Down left		Down right		Up left		Up right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_{V}	50.8	10.5	33.4	5.4	65.6	3.8	61.2	1.9
LTL_{V}	62.6	17.8	41.2	5.4	59.8	2.4	53.6	2.1
UTL_{H}	44.4	2.5	50.8	3.8	55.4	2.4	54.4	2.3
LTL_{H}	57.2	3.9	46.8	5.3	58.8	4.3	53.4	3.2
PAP_{V}	97	38.3	102.4	9.6	12.6	1.1	10.6	0.9
VAP_{V}	30.2	32.9	15.8	1.6	121.4	11.7	107	13.6
PAP_{H}	82.6	59.9	13.4	1.3	116.6	11.1	15.4	1.7
VAP_{H}	24.6	34.3	105.4	8.6	18.8	2.3	99.2	7.9

Table 3. Mean values and standard deviation values of 8 features for subject 2

	Down		Up		Left		Right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_{V}	53.6	3.1	84.8	13.1	16.6	6.8	7.6	4.7
LTL_{V}	73.6	11.8	66.8	7.5	12.8	13.3	3	6.7
UTL_{H}	2.4	4.8	7.2	8.2	52.2	9.4	68.4	5.3
LTL_{H}	-	-	9.2	10.4	69.4	11.5	61.4	9.9
PAP_{V}	183	20.1	19.4	11.1	24.2	28.4	124.6	72.4
VAP_{V}	18.6	3.4	183.8	26.2	47.2	72.1	3.4	7.6
PAP_{H}	9.2	13.1	55.6	93.4	204.4	29	33.4	35.7
VAP_{H}	-	-	80.4	79.3	32.8	28.1	172.6	35.9
	Down left		Down right		Up left		Up right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_{V}	46.8	5.4	41.6	3.6	61.8	2.2	61.8	8.5
LTL_{V}	64.4	3.4	49.6	4.6	50.6	10.8	52	7.2
UTL_{H}	45.2	3.7	65.6	6.6	46.8	9	69.8	7.3
LTL_{H}	46.6	7.6	54.4	13.4	63.6	5.1	67	4.5
PAP_{V}	165.6	22.5	169.8	63.9	22.6	17.1	14	2.5
VAP_{V}	20.6	5	20.2	2.8	209	18.8	188.8	13.7
PAP_{H}	172.4	23.5	54.4	82.5	209.4	18.4	17.8	1.5
VAP_{H}	13.6	2.4	203.6	22.7	26	18	187.8	12.9

from- the vertical and horizontal channels on the middle and bottom panels, respectively. The detection accuracy of eight eye movements is 100% resulting from three healthy subjects. However, the accuracy from other publications is less than 100%. Examples from papers showing the results from four directional eye movements are as follows.

Table 4. Mean values and standard deviation values of 8 features for subject 3

	Down		Up		Left		Right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_V	43.8	8.9	51.6	1.3	4.4	6.1	1.6	3.6
LTL_V	53	3.7	50	8.8	-	-	-	-
UTL_H	1.2	2.7	3.2	7.2	57.6	4.6	66.4	5.5
LTL_H	-	-	2.4	5.4	58.6	3.2	65	3.6
PAP_V	100.6	12.3	30	45.8	21	43.2	3.2	7.2
VAP_V	15.4	1.1	159.2	46.4	-	-	-	-
PAP_H	1.6	3.6	24.6	55	109.8	19.7	14.8	0.4
VAP_H	-	-	44.4	99.3	13.4	0.9	117.2	15.1
	Down left		Down right		Up left		Up right	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std
UTL_V	48.2	7.5	33	4.4	45.8	1.3	57.8	17.3
LTL_V	56.8	2.2	51	10.3	49.8	8.6	57.6	6.6
UTL_H	51.6	4	53.6	2.9	54	4.2	61.6	5.5
LTL_H	61.6	5.4	40	21.4	62	2.3	55.8	5.6
PAP_V	83.8	8.4	127.6	71.9	24.4	30.6	11.2	4.1
VAP_V	18.6	1.1	43.8	61.1	121	34.9	129	17.5
PAP_H	88	7.4	44	59.9	121	34.7	18.8	2.5
VAP_H	14.6	2.4	134.2	68.3	27.2	29	128.2	18.9

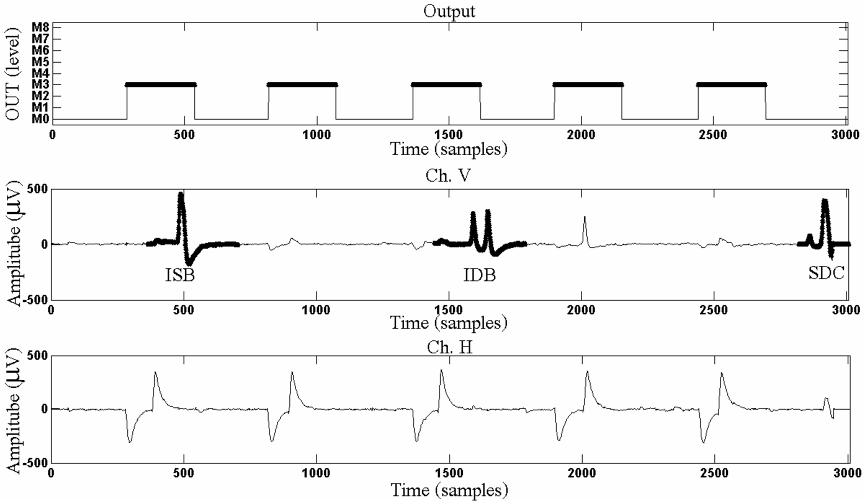


Fig. 5. Effect of noises on the vertical EOG signal: involuntary single blink, involuntary double blinks, and short-duration eye closed

In a work by Deng et al. [16], 90% detection accuracy is achieved for the applications in game control, eye test, and TV controller. In Merino et al. [17], 94% average rate is achieved when the derivative and amplitude levels are used for detecting the direction. Examples from papers showing the results from eight directional eye movements are as follows. In a work by Yamagishi et al. [18], 90.4%

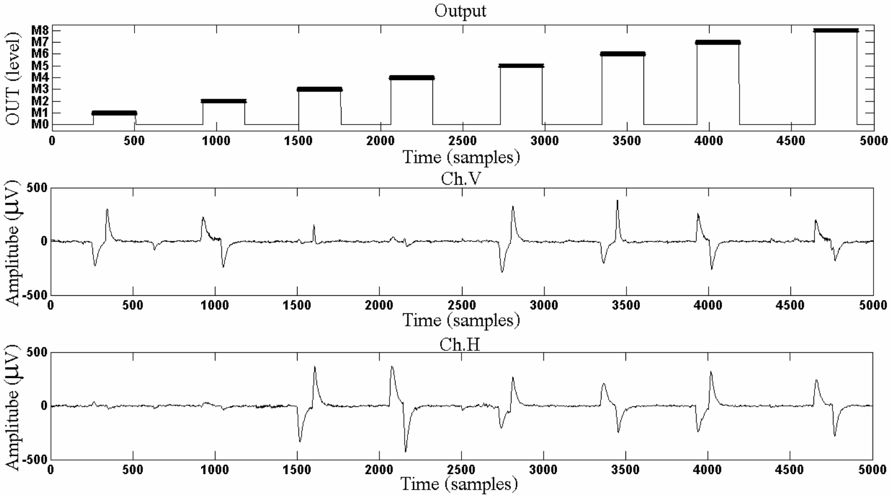


Fig. 6. Example result of EOG classification algorithm for 8 directional movements

accuracy is achieved for the applications in screen keyboard when the algorithm based on logical combination is realized. In Itakura and Sakamoto [19], 96.7% accuracy is obtained from the algorithm based on the integration method when EOG data were acquired from six subjects.

4 Conclusion

EOG signal is widely employed in many clinical applications, such as, evaluation of eye injuries and diagnosis of eye diseases and in many engineering applications, such as, eye-controlled wheelchair and eye-writing recognition. In this paper, we proposed a non-pattern recognition algorithm to classify eight directional movements from EOG signals. From experimental results, the proposed features and threshold analysis showed the best performance to be used in classification of EOG signal. Moreover, the avoiding artifact method that is defined from the first derivative can be effectively used to avoid most noises in EMG signal.

Acknowledgments

This work was supported in part by NECTEC-PSU Center of Excellence for Rehabilitation Engineering, Faculty of Engineering, Prince of Songkla University through contact no. ENG540638S, and by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0110/2550).

References

1. Jia, P., Hu, H., Lu, T., Yuan, K.: Head Gesture Recognition for Hand-free Control of an Intelligent Wheelchair. Int. J. Ind. Rob. 34, 60–68 (2007)
2. Levine, S.P., Bell, D.A., Jaros, L.A., Simpson, R.C., Koren, Y., Borenstein, J.: The NavChair Assistive Wheelchair Navigation System. IEEE Trans. Rehabil. Eng. 7, 443–451 (1999)
3. Evans, D.G., Drew, R., Blenkhorn, P.: Controlling Mouse Pointer Position Using an Infrared Head-operated Joystick. IEEE Trans. Rehabil. Eng. 8, 107–117 (2000)
4. Schmeisser, G., Seamone, W.: An Assistive Equipment Controller for Quadriplegics. Johns Hopkins Med. J. 145, 84–88 (1979)
5. Coyle, E.D.: Electronic Wheelchair Controller Designed for Operation by Hand-operated Joystick, Ultrasonic Non-contact Head Control and Utterance from a Small Word-command Vocabulary. In: IEE Colloquium on New Developments in Electric Vehicles for Disabled Persons, pp. 3/1–3/4. IEEE Press, New York (1995)
6. Tanaka, K., Matsunaga, K., Wang, H.O.: Electroencephalogram-based Control of an Electric Wheelchair. IEEE Trans. Rob. 21, 762–766 (2005)
7. Barea, R., Boquete, L., Mazo, M., Lopez, E.: System for Assisted Mobility Using Eye Movements based on Electrooculography. IEEE Trans. Neural. Syst. Rehabil. Eng. 4, 209–218 (2002)
8. Kim, Y., Doh, N.L., Youm, Y., Chung, W.K.: Robust Discrimination Method of the Electrooculogram Signals for Human-computer Interaction Controlling Mobile Robot. Intell. Autom. Soft Comp. 13, 319–336 (2007)
9. Norris, G., Wilson, E.: The Eye Mouse, an Eye Communication Device. In: 23rd Northeast IEEE Bioengineering Conference, pp. 66–67. IEEE Press, New York (1997)
10. Tsai, J.Z., Lee, C.K., Wu, C.M., Wu, J.J., Kao, K.P.: A Feasibility Study of an Eye-writing System Based on Electro-oculography. J. Med. Biol. Eng. 28, 39–46 (2008)
11. Bulling, A., Ward, J.A., Gellersen, H., Tröster, G.: Eye Movement Analysis for Activity Recognition Using Electrooculography. IEEE Trans. Pattern Anal. Mach. Intell. 33, 741–753 (2011)
12. Measurement of Eye Movement Using Electro Oculography , `http://ee.ucd.ie/~smeredith/EOG_Frameset.htm`
13. Brown, M., Marmor, M., Vaegan, Z.E., Brigell, M., Bach, M.: ISCEV Standard for Clinical Electro-oculography (EOG). Doc. Ophthalmol. 11, 205–212 (2006)
14. Güven, A., Kara, S.: Classification of Electro-oculogram Signals Using Artificial Neural Network. Expert Syst. Appl. 31, 199–205 (2006)
15. Shuyan, H., Gangtie, Z.: Driver Drowsiness Detection with Eyelid related Parameters by Support Vector Machine. Expert Syst. Appl. 36, 7651–7658 (2009)
16. Deng, L.Y., Hsu, C.L., Lin, T.Z., Tuan, J.S., Chang, S.M.: EOG-based Human-Computer Interface System Development. Expert Syst. Appl. 37, 333–3343 (2010)
17. Merino, M., Rivera, O., Gomez, I., Molina, A., Doronzoro, E.: A Method of EOG Signal Processing to Detect the Direction of Eye Movements. In: 1st International Conference on Sensor Device Technologies and Applications, pp. 100–105. IEEE Press, New York (2010)
18. Yamagishi, K., Hori, J., Miyakawa, M.: Development of EOG-Based Communication System Controlled by Eight-Directional Eye Movements. In: 28th EMBS ANNual International Conference, pp. 2574–2577. IEEE Press, New York (2006)
19. Itakura, N., Sakamoto, K.: A New Method for Calculating Eye Movement Displacement from AC Coupled Electro-oculographic Signals in Head Mounted Eye-gaze input Interfaces. Biomed. Signal Process. Control 5, 142–146 (2010)

Recognizing Patterns of Music Signals to Songs Classification Using Modified AIS-Based Classifier

Noor Azilah Draman, Sabrina Ahmad, and Azah Kamilah Muda

Faculty of Information and Communication Technology
University of Technical Malaysia Melaka (UteM)
{azilah,sabrinaahmad,azah}@utem.edu.my

Abstract. Human capabilities of recognizing different type of music and grouping them into categories of genre are so remarkable that experts in music can perform such classification using their hearing senses and logical judgment. For decades now, the scientific community were involved in research to automate the human process of recognizing genre of songs. These efforts would normally imitate the human method of recognizing the music by considering every essential component of the songs from artist voice, melody of the music through to the type of instruments used. As a result, various approaches or mechanisms are introduced and developed to automate the classification process. The results of these studies so far have been remarkable yet can still be improved. The aim of this research is to investigate Artificial Immune System (AIS) domain by focusing on the modified AIS-based classifier to solve this problem where the focuses are the censoring and monitoring modules. In this highlight, stages of music recognition are emphasized where feature extraction, feature selection, and feature classification processes are explained. Comparison of performances between proposed classifier and WEKA application is discussed.

Keywords: Artificial Immune System, modified AIS-based classifier, censoring and monitoring modules, classification, song genre.

1 Introduction

Audio signals contain a great deal of information that can be used to index and classify audio data, particularly music which has led to the consideration of audio classification studies as an important and challenging research area [1]. An efficient mechanism of representing audio data should be used to represent low-level sound properties for describing, recognizing and identifying particular sounds. According to [2], the extracted features used in the classification process need to be comprehensive in which it can represent music data very well; compact where they require small storage space; and efficient in the sense that it can be computed efficiently.

Apart from using the appropriate music features, we also need to use good classifier to classify various categories of music. Since many years ago, many focuses of music related studies are whether to introduce new music features to represent certain aspect of music which often related to introducing new extraction technique to

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 724–737, 2010.

obtain the music features [3], [4] or to manipulate fusion of music features to classify music genre [5].

This research is about investigating on the music signals to search for patterns or features that can be used to recognize and classify music genres. The research is also investigating an algorithm in the AIS domain which focuses on the negative selection algorithm, and proposes a modified version of the algorithm. Previously, there have been a few studies that focused and investigated an approach in the AIS domain called the clonal selection algorithm [6], [7] but there is no other research that concentrates on negative selection algorithm (NSA) before. NSA is applied in this music genre classification study because it has been repeatedly used in pattern recognition studies and produced high quality results. The ability to recognize different patterns using censoring and monitoring modules are also inspired us to investigate the technique and find the solution to the music genre classification problems.

2 Background of Research

In the music genre identification and classification studies, research is initiated to solve problems that occur during recognition such as, deciding which song belongs to which genre. [10], for example has done the early work of classifying songs into different categories of genre using human auditory skills. Finding solutions to increase the performance of the automation process in the classification study is the problem often investigated in the music analysis area. Various attempts to solve this problem have been recorded in [8] – [16]. Not only the problem of automating the process of classification but the question of how to fill the gap of accuracy behind human skilled classification [10] also need to be answered and solved. [3] introduced a new technique to extract the music features called Daubechies Wavelet Coefficient Histograms (DWCHs) with a purpose to overcome the problem of classification accuracies in the previous study. The authors used the Daubechies wavelet filter, *Db8*, to decompose music signals into layers where at the end of each layer they constructed histograms of coefficient wavelet. During experiments they combined the new feature with [3] features and improved the obtained accuracies but not by much.

There is also another attempt that emphasizes on using the pitch, rhythm, and timbre contents to classify music into different genres [17]. [18] proposed a solution to the problem where the authors introduced a new feature extraction method called *InMAF*. [15] attempted to classify the music genre using MIDI (Musical Instrument Digital Interface) and audio features like the pitch, rhythm and timbre features by using the data from [19] study which contained two different sets of contents, the first are MIDI features and the other group are the audio features.

A recent study proposed a new approach to classify music genre by emphasizing on the features from cepstral contents: MFCCs, OSC and MPEG 7 representations [20], where they introduced a novel set features derived from modulation spectral analysis of the spectral representations. [21] developed an AIS-based clustering and classification algorithm which emphasized the ability of the algorithm to adapt and cope efficiently in the complex and changing environments of the immune system. The vital feature of this algorithm compared to other classifiers is their ability to discriminate the self or non-self cells, especially in the situation where the size of non-self cells is larger than self-cells. Later in 2008, they proposed a novel AIS-based classification approach to classify music genres.

[6] proposed a new version of the artificial immune recognition system (AIRS) which was initially developed as a new classification technique based on the humans' immune system [22]. The previous problem of classifying more than two classes at one time has been resolved in this study as ten music genres were classified and produced a better performance with a high accuracy (88 percent). The new version of AIRS has highlighted the nonlinear coefficient of the clonal rate, which has assigned more resources to the detectors with a higher affinity level and has allocated less resource to the detectors with a lower affinity level, which was essential to the recognition processes. The features selection technique applied in the study also contributed to better performances in the music genre classification studies.

[7] described similar experiments to those discussed earlier, where the new version of AIRS classifier has changed the linear method to a nonlinear method of allocating the resources to the clonal rate. Not only did the classifier classify more than two types of cells at one time, the classification performances also produced better performances and provided better accuracies than the previous studies.

3 AIS-Based Music Genre Classification Framework

The artificial immune system (AIS) is defined as mechanisms that manipulate, classify and represent data, and intelligent methodologies that follow a biological paradigm which is the human immune system. It is also an adaptive system that follows the immune theories and functions, principles and method to solve real world problem [23][24]. The definition not only stresses the immune system approach to solve problems, but it also includes mathematical functions that define all the mechanisms of the immune system to be applied to tasks in various fields ranging from optimization, pattern recognition, machine learning, data mining, computer security to fault diagnosis [24]. According to [25], there are at least eleven key elements in the immune system that provide important aspects for the field of information processing: recognizing, extracting features, variety, learning, remembrance, scattered detection, self-regulation, threshold mechanisms, co-stimulation, dynamic protection and probabilistic detection.

AIS are adaptive systems, emulating human body immunology system to solve problems. It is concerned with abstracting the whole concept of immune system to computational systems in solving problems from mathematics, engineering and information technology point of view. AIS is created based upon a set of general purpose algorithms that are modelled to generate artificial components of the human immune system [26]. AIS, is defined as an adaptive system which is enthused by biological immunology and observed functions, principles and models to problem solving.

Negative selection algorithm is introduced in [27] where the idea was inspired by negative selection of T-cells in thymus. The algorithm focused on recognizing self or non-self cells where it will eliminate the T-cells that are not recognized by the thymus. Detail explanations of how negative selection algorithm works can be found in [28]. As has been investigated before, it would be impossible to apply NSA without modification in each study as each problem and solutions were also different. However, we will not discuss the NSA and how it changes in each study as it is not in this research scope. To be able to apply the AIS approach in solving the music genre recognition problem, we need to follow and understand the basic things that are needed based on the basic elements of AIS framework [23] [29]:

a) *a representation for the components of the system*
b) *a set of mechanisms to evaluate the interaction of elements with the environment and with each other*
c) *procedures of adaptation that govern the dynamic of the system*

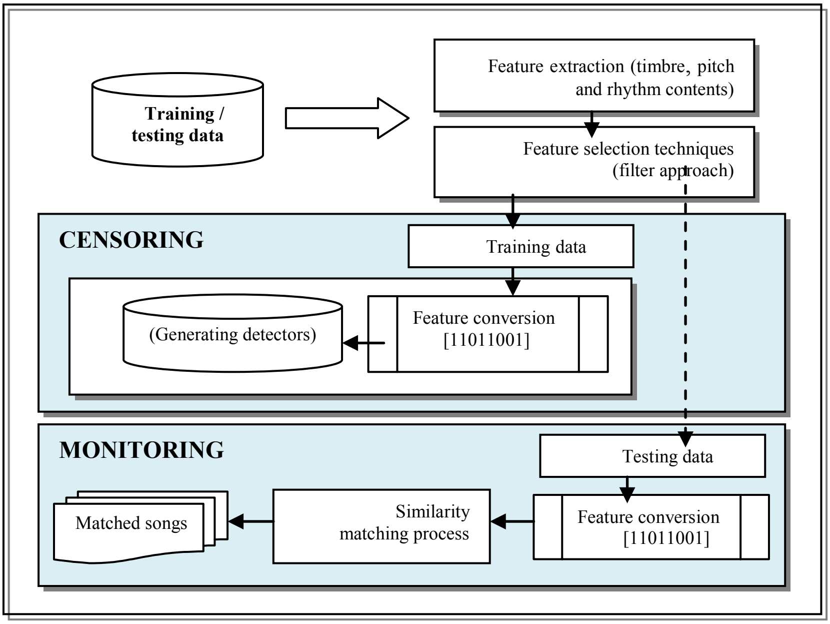

Training / testing data
Feature extraction (timbre, pitch and rhythm contents)
Feature selection techniques (filter approach)
CENSORING
Training data
(Generating detectors)
Feature conversion [11011001]
MONITORING
Testing data
Feature conversion [11011001]
Similarity matching process
Matched songs

Fig. 1. AIS-based music genre classification framework

Figure 1 illustrates the structures of our proposed AIS-based music genre classification framework. We begin the process by extracting music features from the data before we transform them into binary bit strings. Two highlighted processes, the censoring and monitoring modules depict two important parts of AIS-based model in recognizing different patterns. The artificial immune system (AIS) is defined as mechanisms that manipulate, classify and represent data, and intelligent methodologies that follow a biological paradigm which is the human immune system. It is also an adaptive system that follows the immune theories and functions, principles and methods to solve real world problems [23][29].

The following sections discuss each step in the recognition stage, that are the feature extraction, feature selection and feature classification which comprises two important processes; censoring and monitoring.

3.1 Music Features Extraction

The feature extraction in the content-based definition is a process of calculating the music contents to provide numerical representations to characterize the music [3]. In order to get quality feature representations and quality results in the music recognition work, the choice of selected features should be able to reflect the underlying information about the music. For example, if there are different genres of music that need to be identified, the extracted music features should be able to represent all types of those genres. According to [30], the extracted features from music data should be able to fulfil two criteria. The first criterion is related to the feature space where objects that are considered similar will be located next to each other and the differences between object regions can be clearly seen on that feature space. The second criterion is the extracted technique used should be able to conserve all-important information contained in the data. The human perception of the sounds is the way listeners generate music label during the identification process and it happens whenever they hear the sounds. Technically, the humans also depend on the features extracted from the sounds they hear in order to recognize the sounds. The auditory system in human being is a very sophisticated system where it can automatically separate the sounds and immediately identify the source.

[30] stated there are two categories of the extracted features used in music signal analysis: 1) perceptual features and 2) physical features. The perceptual features were based on the perception of sounds when humans hear them, for example the high or low pitch, the melody and the frequency. The physical features on the other hand were produced in mathematical computing mechanisms during the sounds analysis using signal-processing concepts. The perceptual and physical features were related to each other. The physical features were labelled as physical because the extraction process imitated the perceptual features that human processed, for example, a slow song that contains harmony sounds is assumed as a Ballard song because the physical features used to recognize it were calculated from the song's slow beat using mathematical functions. Three music contents were introduced in [3], which are the pitch, timbre, and rhythm. Further elaboration on the music contents are as followed:-

Timbre. In music, timbre is defined as the quality of sounds or the colours of music and is produced when a musical instrument played music notes that contained more than one level of frequencies. It allows a person to distinguish different instruments playing the same pitch and loudness. The human ear and brain have magnificent capabilities that can detect even a very small variation of timbre quality in the music. These capabilities were the reason why humans can discriminate two different instruments playing similar notes, similar pitch, or loudness.

The major aims of the timbre perception studies are to develop a theory of sounds classification [32]. In the study, the author focused on experimenting different sounds coming from various orchestra instruments. The sounds contained similar pitch, loudness, and duration. The study aimed to get perceptual relationships between the instruments and it showed that timbre could discriminate different sources of sounds in a note. Researchers and experts agree that timbre quality main function in music signals is to recognize different instruments playing a note with similar pitch and loudness within similar duration. According to [34], timbre content is the common music feature that is used to distinguish different aspects of the music and instrumentations, and if they are combined with other features such as rhythm and tonal characteristics, they usually are enough to discriminate various styles of the music.

Rhythm. Rhythm, by musical definition, is the musical time. It relates to the beat of music. Beat represents the music notes that can be in a whole, a half or a quarter long. One whole note represents a length of four beats. Rhythm is the flow of music. It organizes the notes within the music pace and tempo is the term used to indicate the arrangement of the music pace. Tempo and rhythm is normally used together to describe a slow or a fast song.

In music analysis research, many studies have focused on the rhythm content that emphasized the tempo and beat of the music to recognize the songs. Rhythm can be represented by various terms in music analysis, ranging from low level audio signal features to a more abstract or symbolic concept [33]. Theoretically, the rhythm represents the beat, and the beat is the note of the music. Note in general, represents certain patterns of a song and these patterns technically can be used to recognize the music. The music pattern is the symbolic information that represents various types of music for example, songs from different regions or ethnic groups that generally tell stories about their way of life or anything related to their community through their songs.

There are studies that focused on analysing repeating patterns of the music notes to retrieve songs [34], and introducing new features from music objects to recognize music information such as music themes [35]. In their work, [35] emphasized the notes to locate patterns that reappeared more than once in a song and they agreed with [33] descriptions about the rhythm contents which can be used symbolically to recognize a certain pattern of a song. This can be very useful in the music identification process especially songs that can be recognized using the themes or stories.

Pitch. In music, pitch is normally associated with a high or a low tone, which depends on the frequencies or vibration rates of a music sound. A frequency is the number of vibrations per second and is measured in Hertz (Hz). A high frequency means high pitch and a low frequency means low pitch. The high or low pitch of a tone in a sound note is the listener's evaluation of the frequencies. Two different methods of extracting pitch contents were presented, where one focused on the phrase-based melody extraction [36] whereas the other used mathematical computations to extract the contents from the music signals [3] [37] [38][39]. In the first method, [36] focused on relative pitch sequences that were obtained by converting the song notes to different levels of pitch, which is higher than, or equal to, or lower than the previous note.

The second method used a computationally mathematical model to extract the pitch contents from the complex audio signals. The computation involves an autocorrelation tool that is useful to find repeating patterns in a signal. This tool is used to determine the presence of a periodic signal that contained noise and to identify the fundamental frequency of a signal. [38] was the first to mention the proposed method, called auditory model to compute a multi-pitch analysis model considering an auditory modelling point of view.

A recent study that adopted the auditory model is in [39] where the author compared the pitch analysis content using an auditory model and conventional methods of obtaining the pitch contents. The study showed that the auditory model has greater advantages than the conventional techniques in obtaining the pitch sequences using the frequency-related analysis. Following the work of [38], [37] applied similar technique to extract the pitch contents from the music signals.

3.2 Music Feature Selection

The music representations used a wide set of extracted features from various music contents, such as timbre, pitch and rhythm. Among the features, some are irrelevant and redundant for music recognition processes. These irrelevant and redundant features need to be eliminated before we use the rest of the features in recognition processes. The reason for the elimination is that, music recognition can improve the classification performances and shorten the processing [40]. Selecting the relevant features from a wide range of extracted features is a challenging work [41]. The term relevant as applied in the literature normally depends on the question of relating the relevancy of features to something else [42].

3.3 Music Genre Classification

In this stage, we introduced the Modified AIS-Based Classifier that contained two important modules of Negative Selection Algorithm (NSA). the censoring and monitoring. Censoring is a process where detectors are generated, and monitoring is a process where comparison between detectors and antigen are made to find the match. Modified AIS-based classifier is proposed after some modifications and adjustments are applied to the NSA. These works are important in our research as it is to enable the proposed classifier to solve the music genre classification problem. Figure 2 shows the building blocks of the proposed classifier in this last stage of the classification framework where three important mechanisms of NSA is illustrated, which are the binary bit string conversion, censoring and monitoring modules

Censoring. This module is described as a module that produces detectors, which is the key aspect of identification. Censoring module normally starts after feature extraction and feature selection finished. It involves data features conversion where the features will be represented by binary bit strings (for example, feature vector = -3.4523123 is converted using –XOR operation and becomes 101011001). After the conversion, the binary bit strings will then go through the complementary process and become the detectors.

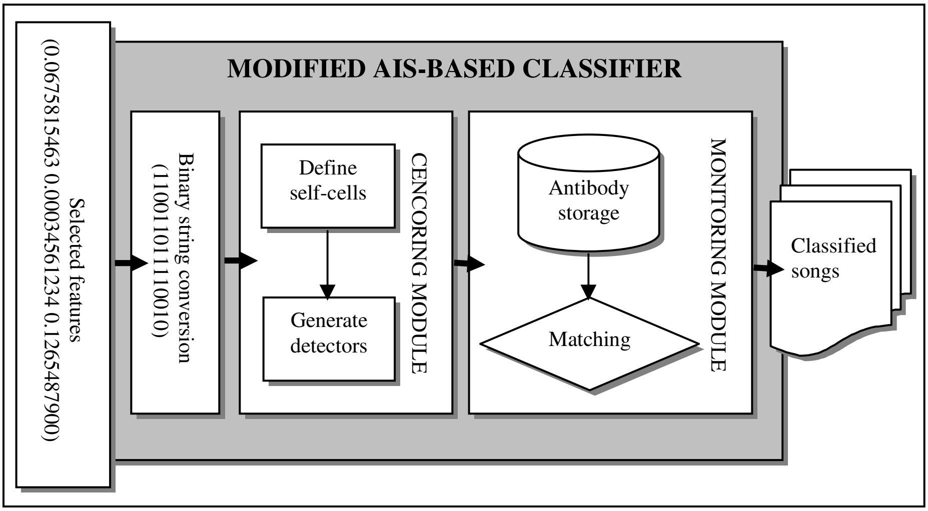


Fig. 2. Stages of classification task involving censoring and monitoring modules

The detectors are created based on how many song genres that are needed to be classified. During the process, antigens are compared with the detector candidates to evaluate the affinity binding (similarity values). The process applies the XOR operation to determine and evaluate the affinity binding between them. The threshold value is used as a benchmark in the process. As each antibody is consisting of 15 binary digits, the threshold value is set to a certain value in order to evaluate the affinity binding between detectors and antigen (song genres). In the XOR operation, values "0" and "1" are used and are counted to decide whether the matched bits exceed the threshold value or not. As the algorithm considers the non-self cells as detectors, the not match antigen-detector will be based on the "1" value. The higher the "1" than "0" value during comparison, more non-self cells are indicated. Once identified, the cell then is considered as a detector and is stored for further analysis in the next module. The detectors by using the training data where in the process, a training model will be created and used in the identification and classification processes.

Monitoring. This module starts once the detectors (training model) are created and are compared to the antigens (we used testing data as antigens) to find similarity between these data and calculate the affinity binding. The comparison is referring to the classification task and when it produces binary bit '1', the data is considered bind. However, in this scenario, we will use the word 'match' instead of 'bind' to define the similarities. In the original version of NSA, the process uses value "0" to indicate similarity or 'match' during monitoring process. The more of '0' found, the more similar the antigen to the detector. Once matched, the antigen is considered as self cell and will be ignored. Since the objective of NSA is to identify non-self cells to recognize antigens, the 'non-match' cells are detected and a change situation is assumed occurred in the network security.

The comparison of value ‘0’ is simple and straightforward however, according to[43], the term ‘match’ as used in the original version of NSA did not give any specific meaning, it is too general, and did not specify the type of representation space used.

Another important factor in the monitoring module is the threshold value. The value is used to set the benchmark number of binary bits that both antigen and detector cells should bind, because it will decide whether they are matched or not. Both cells are considered matched if the bind bits are exceeding the threshold value. The value used in the experiments generally indicates the reliability levels of the results where the higher the value used means reliable results are obtained from the similarity matching process.

Classification accuracy. In the proposed modified AIS-based classifier, we combined all feature vectors from the music contents (pitch, rhythm, and timbre contents). Table 1 discusses the computation stages, where the first stage of calculation is applied to identify and compute the bits between both cells that are matched and then get the match percentage. In the next stage, the calculation is to get the threshold value percentage where the value will be the indicator used to decide whether each dataset is matched or not. The last stage of calculation is to get the classification accuracy where all matched song are divided with the amount of total tested data and then get the percentage.

Table 1. Proposed classification accuracy method

Category	Calculation formulas
Data genre accuracy stage	Σ bits_matched / Σ features_bits x 100
Threshold (*r*) %	(Σ *r* * num_of_features / Σ bits_per_feature * num_of_features) x 100
Dataset accuracy stage	(Num_of_genre_match / num_of_testing_data) x 100

Four binary matching techniques are applied in the AIS, which are the Hamming distance matching, the r-chunk matching, the r-contiguous matching and the multiple r-contiguous matching rules.

4 Experimental Results

The songs that we used in our experimental work comprises of ten different genres in the Western song collections, which are Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. One thousand songs (courtesy of MARSYAS group research) are used in the experiments. Two applications were used to extract these features, which are MARSYAS [44] and rhythm pattern extraction tool [45]. We prepared training and testing datasets where similar data is used in the experiments except the data is in the attribute related file format (ARFF) for WEKA experiments and in the data file (DAT) format for modified AIS-based classifier demonstrations. Two attribute evaluators are the CFSSubsetEval and the ConsistencySubsetEval which apply both BestFirst and GreedyStepwise search are used to select significant features in the experiments. The selected features are tested separately in the classification experiments to find which significant selected features produce the highest classification accuracy.

We evaluate the proposed AIS-based music genre classification algorithm based on the results produced in the classification experiments using our proposed modified AIS-based classifier and WEKA application. We have conducted the experiments using classifiers from WEKA application such as the k-nearest neighbour learner, decision tree, support vector machine and naive-bayes. The tested classifiers are the bayes-net, sequential minimal optimisation (SMO), IB1, J48 and Bagging.

Two setup cases of experiments were prepared, which are according to the binary similarity matching techniques, and the feature selection techniques. The classification performances are evaluated according to the classifiers used in the experiments. The music contents are individually classified in the classification evaluation. From the three contents, the music features that we have extracted are categorized into five main groups:

1) timbre-related features consisting MFCC, zero-crossings rate, spectral centroid, spectral flux, and spectral roll-off,
2) chroma-related features consisting 12 music notes (A, A#, B, C, C#, D, D#, E, F, F#, G and G#),
3) rhythm pattern (RP),
4) rhythm histogram (RH), and
5) statistical spectrum descriptor (SSD) features

The following figures illustrate the classification results according to the setup cases.

Figure 3 and 4 illustrate the performance of classification experiments using various similarity matching techniques (R-Chunk (RCH), Hamming Distance (HD), R-Contiguous (RCo) and Multiple R-Contiguous (MRCo)) in the modified AIS-based

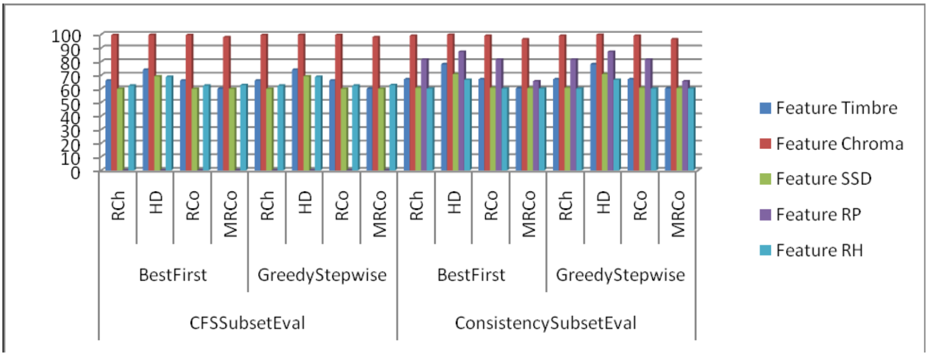


Fig. 3. The classification performances of modified AIS-based classifier using different binary matching techniques

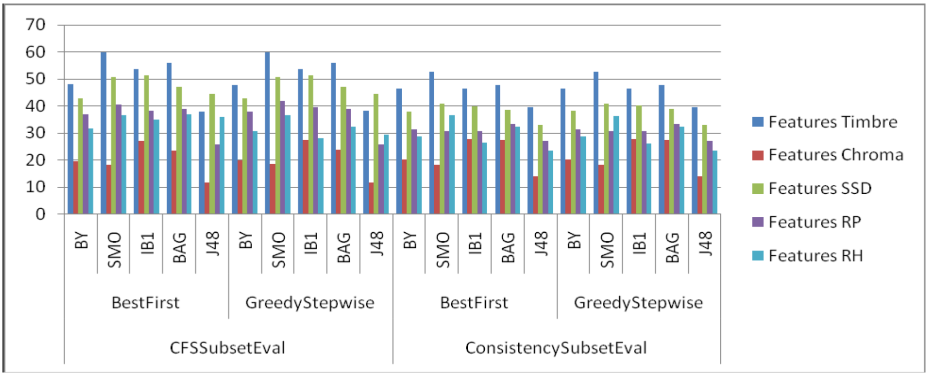


Fig. 4. The classification performances of WEKA classifiers

algorithm and classifiers in WEKA application. Both feature selection techniques, CFSSubsetEval and ConsistencySubsetEval are compared in the performance evaluation. Overall, we can see that the results from the proposed classifier averagely are higher than WEKA classifiers by 20 – 30 percents. Among the matching techniques, HD technique has consistently produced classification accuracies between 70 – 90 percent when evaluated with the feature vectors selected using the Consistency SubsetEval technique.

5 Conclusions

The availability of techniques and methods for classification in music analysis field today has shown that researchers in this area are very concerned with the performance. As the collections of digital songs keep increasing online, their studies have contributed a major breakthrough to the internet users and others.

In this paper, we have explained and evaluated the proposed modified AIS-based classifier in different category of experiments. In each experiment, the proposed classifier outperformed any performance from other classifiers. The classification results clearly show that the proposed modified AIS-based classifier is a new algorithm or mechanism to solve problem in the area of music genre classification.

We strongly believe that our discussion throughout this paper has given opportunities to other researchers in this area of studies to fill the gaps, to explore further and to provide solutions to the known and un-known problem that has yet to be discovered. Future work will include an investigation on how to manage the threshold value efficiently and probably, exhaustive search approach should be applied to evaluate the highest threshold value that can provide high classification accuracies.

References

1. Kim, H.G., Moreau, N., Sikora, T.: Audio classification based on MPEG-7 spectral basis representations. IEEE Transactions on Circuits and Systems for Video Technology (2004b)
2. Li, T., Ogihara, M.: Toward intelligent music information retrieval. IEEE Transactions on Multimedia 8, 564–574 (2006)

3. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10(5), 293–302 (2002)
4. Li, T., Ogihara, M., Zhu, S.H.: Integrating features from different sources for music information retrieval. In: Perner, P. (ed.) ICDM 2006. LNCS (LNAI), vol. 4065, pp. 372–381. Springer, Heidelberg (2006)
5. SillaA, C.N., KoerichH, A.L., Kaestner, C.A.A.: Improving automatic music genre classification with hybrid content-based feature vectors. In: 25th Symposium on Applied Computing, Sierre, Switzerland (2010)
6. Golzari, S., Doraisamy, S., Sulaiman, M.N., Udzir, N.I.: Hybrid Approach to Traditional Malay Music Genre Classification: Combining Feature Selection and Artificial Immune Recognition System. In: Proceedings International Symposium of Information Technology 2008, vol. 1-4, pp. 1068–1073 (2008a)
7. Golzari, S., Doraisamy, S., Sulaiman, M.N.B., Udzir, N.I., Norowi, N.M.: Artificial immune recognition system with nonlinear resource allocation method and application to traditional malay music genre classification. In: Bentley, P.J., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS, vol. 5132, pp. 132–141. Springer, Heidelberg (2008)
8. Draman, A.K.: Authorship invarianceness for writer identification using invariant discretiation and modified immune classifier., PhD thesis. University of Technology Malaysia (2009)
9. Lippens, S., Martens, J.P., Mulder, T.D.: A comparison of human and automatic musical genre classification. Acoustics, Speech, and Signal Processing (2004)
10. Brecheisen, S., Kriegel, H.P., Kunath, P., Pryakhin, A.: Hierarchical genre classification for large music collections. In: Proceedings IEEE International Conference on Multimedia and Expo - ICME 2006, vol. 1-5, pp. 1385–1388 (2006)
11. Ahrendt, P., Larsen, J., Goutte, C.: Co-occurrence models in music genre classification. In: 2005 IEEE Workshop on Machine Learning for Signal Processing (MLSP), pp. 247–252 (2005)
12. Bağcı, U., Erzin, E.: Boosting classifiers for music genre classification. In: Yolum, p., Güngör, T., Gürgen, F., Özturan, C. (eds.) ISCIS 2005. LNCS, vol. 3733, pp. 575–584. Springer, Heidelberg (2005)
13. Bagci, U., Erzin, E.: Inter genre similarity modeling for automatic music genre classification. In: 2006 IEEE 14th Signal Processing and Communications Applications, vol. 1, 2, pp. 639–642 (2006)
14. Cataltepe, Z., Yaslan, Y., Sonmez, A.: Music genre classification using MIDI and audio features. Eurasip Journal on Advances in Signal Processing (2007)
15. Cheng, H.T., Yang, Y.H., Lin, Y.C., Liao, I.B., Chen, H.H.: Automatic Chord Recognition for Music Classification and Retrieval. In: IEEE International Conference on Multimedia and Expo., vol. 1-4, pp. 1505–1508 (2008)
16. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of the 26th Annual International ACM SIGIR., Toronto, Canada, pp. 282–289 (2003)
17. Neumayer, R., Rauber, A.: Integration of text and audio features for genre classification in music information retrieval. In: Proceeding of 29th European Conference on Information Retrieval, Rome, Italy, pp. 724–727 (2007)
18. Shen, J., Shepherd, J.A., Ngu, A.H.H.: On efficient music genre classification. In: Zhou, L.-z., Ooi, B.-C., Meng, X. (eds.) DASFAA 2005. LNCS, vol. 3453, pp. 253–264. Springer, Heidelberg (2005)
19. Mckay, C., Fujinaga, I.: Musical genre classification: Is it worth pursuing and how can it be improved. In: ISMIR 2006, Victoria, Canada (2006)

20. Lee, C.H., Shih, J.L., Yu, K.M., Lin, H.S.: Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features. IEEE Transactions on Multimedia 11, 670–682 (2009)
21. Sotiropaolos, D.N., Lampropaolos, A.S., Tsihrintzis, G.A. (Artificial Immune System-Based Music Genre Classification. New Directions in Intelligent Interactive Multimedia 142, 191–200 (2008)
22. Watkins, A.B.: AIRS: A resource limited artificial immune classifier. In: Computer Science, Mississippi State University, Mississippi (2001)
23. de Casto, L.N., Timmis, J.: Artificial immune system: A new computational intelligence approach, pp. 76–79. Springer, Great Britain (2002)
24. Dasgupta, D.: Information processing mechanisms of the immune system. In: Corne, D., Dorigo, M., Glover, F. (eds.) New Ideas in Optimization, McGraw Hill, London (1999)
25. Xiao, R.-B., Wang, L., Liu, Y.: A framework of AIS based pattern classification and matching for engineering creative design. In: IEEE First International on Machine Learning and Cybernetics, Beijing, China (2002)
26. de Casto, L.N., Timmis, J.: ‘Artificial immune system: A new computational intelligence system: A new Computational Intelligence (2001)
27. Xiao, R.-B., Wang, L., Liu, Y.: A framework of AIS based pattern classification and matching for engineering creative design. In: Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, China, pp. 1554–1558 (2002)
28. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself discrimination in a computer. In: Proceedings of 1994 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, CA, USA, pp. 202–212 (1994)
29. Timmis, J., Andrews, P.S.: L.Owen, N. D. & Clark, E. An interdisciplinary perspective on artificial immune system. Evolutionary Intelligence 1, 5–26 (2008)
30. Kosina, K.: Music genre recognition. Media Technology and Design (MTD). Upper Austria University of Applied Sciences Ltd, Hagenberg (2002)
31. Ellis, D.P.W.: Prediction-driven computational auditory scene analysis for dense sound mix-tures. In: ESCA Workshop on the Auditory Basis of Speech Perception, Keele, UK (1996)
32. Grey, J.M.: Multidimensional perceptual scaling of musical timbres. Acoustical Society of America 61, 1270–1277 (1977)
33. Gouyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhythmic decriptions for musical genre classification. In: AES 25th International Conference, London, UK (2004)
34. Hsu, J.-L., Liu, C.-C., Chen, A.L.P.: Discovering nontrivial repeating patterns in music data. IEEE Transactions on Multimedia 3, 311–325 (2001)
35. Karydis, I., Nanopaolos, A., Manolopoulos, Y.: Finding maximum-length repeating patterns in music databases. Multimedia Tools and Applications 32, 49–71 (2007)
36. Yanase, T., Takasu, A., Adachi, J.: Phrase Based Feature Extraction for Musical Information Retrieval. In: Communications, Computers and Signal Processing, Victoria BC Canada (1999)
37. Tzanetakis, G., Ermolinskyi, A., Cook, P.: Pitch histograms in audio and symbolic music information retrieval. In: ISMIR 2002, Pompidao, Paris (2002)
38. Tolonen, T., Karjalainen, M.: A computationally efficient multipitch analysis model. IEEE Transactions on Speech and Audio Processing 8, 708–716 (2000)
39. Klapuri, A.: Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Transactions on Audio Speech and Language Processing 16 (2008)
40. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering 9, 642–645 (1997)

41. Liu, H., Dougherty, E.R., DY, J.G., Torkolla, K., Tuv, E., Penh, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, L., Forman, G.: Evolving feature selection. IEEE Intelligent Systems 20, 64–76 (2005)
42. Gonzalez, F., Dasgupta, D., Gomez, J.: The effect of binary matching rules in negative selection. In: Cantú-Paz, E., Foster, J.A., Deb, K., Davis, L., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Kendall, G., Wilson, S.W., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A., Dowsland, K.A., Jonoska, N., Miller, J., Standish, R.K. (eds.) GECCO 2003. LNCS, vol. 2724, Springer, Heidelberg (2003)
43. Tzanetakis, G., Cook, P.: MARSYAS: a framework for audio analysis. Organized Sound 4, 169–175 (1999)
44. Lidy, T.: Evaluations of new audio features and their utilization in novel music retrieval applictions. Vienna University of Technology, Vienna (2006)

Author Index