Volume 6 • Issue 4 • August 2012

Editor-in-Chief
Dr. Chen-Chi Shing

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

ISSN : 1985-1553

Publication Frequency: 6 Issues / Year

CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

VOLUME 6, ISSUE 4, 2012

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-1553
International Journal of Computer Science and Security is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCSS Journal is a part of CSC Publishers
Computer Science Journals
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

Book: Volume 6, Issue 4, August 2012
Publishing Date: 10 - 08- 2012
ISSN (Online): 1985 -1553

IJCSS Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2012

EDITORIAL PREFACE

This is fourth issue of volume six of the International Journal of Computer Science and Security (IJCSS). IJCSS is an International refereed journal for publication of current research in computer science and computer security technologies. IJCSS publishes research papers dealing primarily with the technological aspects of computer science in general and computer security in particular. Publications of IJCSS are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJCSS are databases, electronic commerce, multimedia, bioinformatics, signal processing, image processing, access control, computer security, cryptography, communications and data security, etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 6, 2012, IJCSS appears in more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJCSS is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJCSS as one of the top International journal in computer science and security, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Computer science and security fields.

IJCSS editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCSS provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computer Science and Security (IJCSS)

EDITORIAL BOARD

Assistant Professor Vishal Bharti
Maharishi Dayanand University
India

Dr. Parvinder Singh
University of Sc. & Tech
India

Assistant Professor Vishal Bharti
Maharishi Dayanand University,
India

TABLE OF CONTENTS

Volume 6, Issue 4, August 2012

Pages

Data Security In Relational Database Management System

R.Balasubramaniam *bala27788@gmail.com*
Assistant Professor/Department of Computer Science
Kathir College of Engineering
Coimbatore-641062, Tamilnadu, India

Abstract

Proving ownerships rights on outsourced relational database is a crucial issue in today's internet based application environments and in many content distribution applications. Here mechanism is proposed for proof of ownership based on the secure embedding of a robust imperceptible watermark in relational data. Watermarking of relational databases as a constrained optimization problem and discus efficient techniques to solve the optimization problem and to handle the constraints. This watermarking technique is resilient to watermark synchronization errors because it uses a partioning approach that does not require marker tuple. This approach overcomes a major weakness in previously proposed watermarking techniques. Watermark decoding is based on a threshold-based technique characterized by an optimal threshold that minimizes the probability of decoding errors. An implemented a proof of concept implementation of our watermarking technique and showed by experimental results that our technique is resilient to tuple deletion, alteration and insertion attacks.

Keywords: Embedding Algorithm, Watermarking & Threshold Value

1. INTRODUCTION

The rapid growth of the Internet and related technologies has offered an unprecedented ability to access and redistribute digital contents. In such a context, enforcing data ownership is an important requirement, which requires articulated solutions, encompassing technical, organizational, and legal aspects. Although we are still far from such comprehensive solutions, in the last years, watermarking techniques have emerged as an important building block that plays a crucial role in addressing the ownership problem. Such techniques allow the owner of the data to embed an imperceptible watermark into the data. A watermark describes information that can be used to prove the ownership of data such as the owner, origin, or recipient of the content. Secure embedding requires that the embedded watermark must not be easily tampered with, forged, or removed from the watermarked data. Imperceptible embedding means that the presence of the watermark is unnoticeable in the data. Furthermore, the watermark detection is blinded, that is, it neither requires the knowledge of the original data nor the watermark. Watermarking techniques have been developed for video, images, audio, and text data and also for software and natural language text. By contrast, the problem of watermarking relational data has not been given appropriate attention. There are, however, many application contexts for which data represent an important asset, the ownership of which must thus be carefully enforced. This is the case, for example, of weather data, stock market data, power consumption, consumer behavior data, and medical and scientific data. Watermark embedding for relational data is made possible by the fact that real data can very often tolerate a small amount of error without any significant degradation with respect to their usability. For example, when dealing with weather data, changing some daily temperatures of 1 or2 degrees is a modification that leaves the data still usable.[8]

To date, only a few approaches to the problem of watermarking relational data have been proposed. These techniques, however, are not very resilient to watermark attacks. In this paper, we present a watermarking technique for relational data that is highly resilient compared to these techniques. In particular, our proposed technique is resilient to tuple deletion, alteration, and insertion attacks. The main contributions of the paper are summarized as follows: We formulate the watermarking of relational databases as a constrained optimization problem and discuss efficient techniques to handle the constraints.

2. Existing System

Watermarking in least significant bits (LSB).This technique embeds the watermark bits in the least significant bits of selected attributes of a selected subset of tuple's. It uses secret key in watermarking. For each tuple's a secure message, authenticated code is computed using the secret key and tuple's primary key. The computed MAC is used select candidate tuple's attributes and the LSB positions in the selected attributes. This technique does not provide mechanism for multi bit watermarks. The watermark can be easily compromised by very trivial attacks [11].

2.1 Drawbacks

This technique does not provide mechanism for multi bit watermarks. The watermark can be easily compromised by very trivial attacks

3. Proposed System

Watermarking embeds ownership information in digital content. Watermark describes information that can be used to prove the ownership of relational database. Here the embedding is hidden that the presence of watermarking is invisible to the user. It is not resilient to watermark attacks. Optimal threshold reduces probability of decoding error. Multiple embedding of watermark bits in the dataset increases additional security.

3.1 Objective of Proposed work

It is not resilient to watermark attacks. Optimal threshold reduces probability of decoding error. Multiple embedding of watermark bits in the dataset increases additional security.

4. MODULE DESCRIPTION

4.1 SERVER (SOURCE) MODULE

- Main Form has controls like select Table Name and specify Destination System Name

- DBCON –
  This class establishes a connection to the database using SQL SERVER.

- ACTION CONTROLLER –
  This class initiates the process by accessing the data table.

- PARTITION –
  This class eventually divides the table records and assigns partition number to it.

- SINGLE BIT ENCODING –
  This class encodes the partitioned file by adding one bit to each record.

- ENCRYPTION –
  This class gets all the encoded records and encrypts them as a whole file along the secret key.

- WATERMARK SERIALIZATION –
  This class serializes (Convert The Object into file) and sends it to the destination via network connections.

4.2 CLIENT (DESTINATION) MODULE

- MAIN FORM –

This class receives the file from source through thread.

- WATERMARK DESERIALIZE –

This class reads the received file and convert the stream as file.

- DECRYPTION –

This class decrypts the file using the secret key.

- SINGLE BIT DECODING –

This class gets the file as an object from the Hash Table using the key value added to it and also decodes the file by removing the additional bit added to each record.

- REVERSE PARTITION –

This class merge the partitioned file or records into a single file.

- DATA INSERT –

This class reads each record and stores it in Hash Table and which in turn is inserted into the Database one by one using SQL SERVER.

4.3 DATA PARTITIONING

Data partitioning in relational data warehouse can implemented by objects partitioning of base tables, clustered and non-clustered indexes, and index views. Range partitions refer to table partitions which are defined by a customizable range of data. The end user or database administrator can define the partition function with boundary values, partition scheme having file group mappings and table which are mapped to the partition scheme. By using secret key the data set is partitioned into several non overlapping partitions.

4.4 WATERMARK EMBEDDING – SINGLE BIT ENCODING

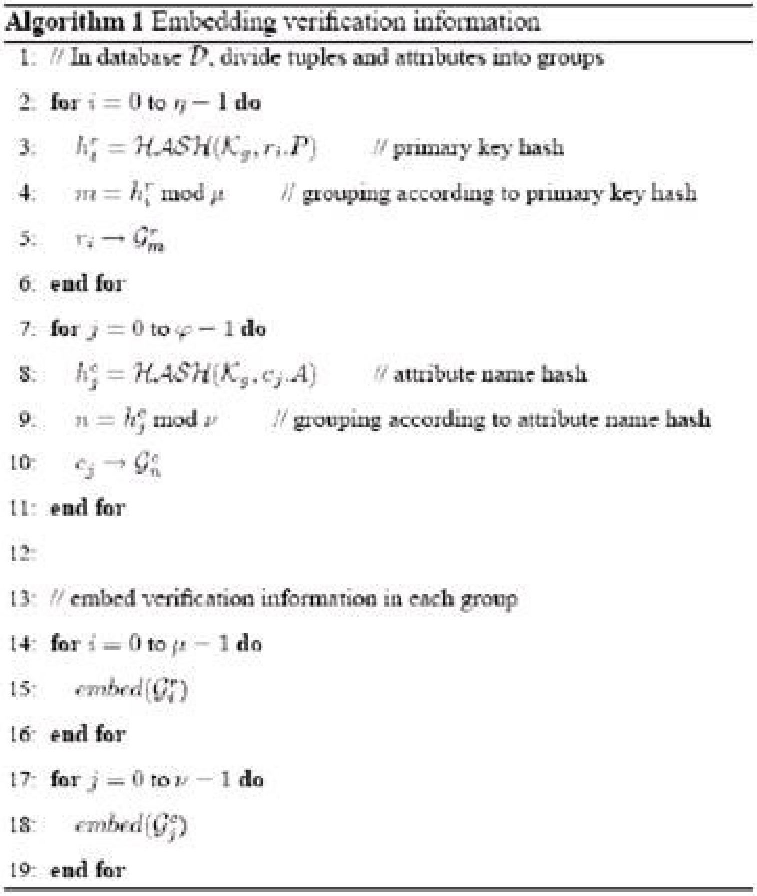
A watermark bit is embedded in each partition by Single Bit Encoding algorithm (figure.1). Watermarking is a technology for embedding various types of information in digital content. In general, information for protecting copyrights and proving the validity of data is embedded as a watermark. Watermarks are added to images or audio data in such a way that they are invisible or inaudible Ñ unidentifiable by human eye or ear. Furthermore, they can be embedded in content with a variety of file formats. Watermarking is the content protection method for the multimedia era.

4.5 OPTIMAL THRESHOLD EVALUATION

The bit embedding statistics are used to compute the optimal threshold that minimizes the probability of decoding error. The optimization technique used in this experiment is pattern search technique (PS). PS methods are direct search methods for non-linear optimization. It starts at an initial point and samples the objective function at a predetermined pattern of points centered about that point with the goal of producing a new better iterate

4.6 THRESHOLD BASED DECODING

The statistics of each partition are evaluated, and the embedded is decoded using a threshold based scheme based on the optimal threshold. The probability of bit decoding error is de.ned as the probability of an embedded bit decoded incorrectly. The decoding threshold T_ is selected such that it minimizes the probability of decoding error. The bit embedding stage is based on the maximization or minimization of the tail count; these optimized hiding function values computed during the encoding stage are used to compute the optimum threshold T.



```
Algorithm 1 Embedding verification information
// In database $\mathcal{D}$, divide tuples and attributes into groups
for $i = 0$ to $\eta - 1$ do
    $h_i^r = \mathcal{HASH}(\mathcal{K}_g, r_i.P)$        // primary key hash
    $m = h_i^r \bmod \mu$        // grouping according to primary key hash
    $r_i \rightarrow \mathcal{G}_m^r$
end for
for $j = 0$ to $\varphi - 1$ do
    $h_j^c = \mathcal{HASH}(\mathcal{K}_g, c_j.A)$        // attribute name hash
    $n = h_j^c \bmod \nu$        // grouping according to attribute name hash
    $c_j \rightarrow \mathcal{G}_n^c$
end for

// embed verification information in each group
for $i = 0$ to $\mu - 1$ do
    $embed(\mathcal{G}_i^r)$
end for
for $j = 0$ to $\nu - 1$ do
    $embed(\mathcal{G}_j^c)$
end for
```

FIGURE 1: Single Bit Encoding Algorithm

4.7 ENTITY RELATIONSHIP DIAGRAM

Entity-Relationship model (ER model for short) is an abstract and conceptual representation of data. Entity-relationship modeling is a database modeling method, used to produce a type of conceptual schema or semantic data model of a system, often a relational database, and its requirements in a top-down fashion. Diagrams created by this process are called entity-relationship diagrams or ER diagrams. The ER diagram for Sender Activity and Receiver Activity is shown in figure.2 & figure 3.

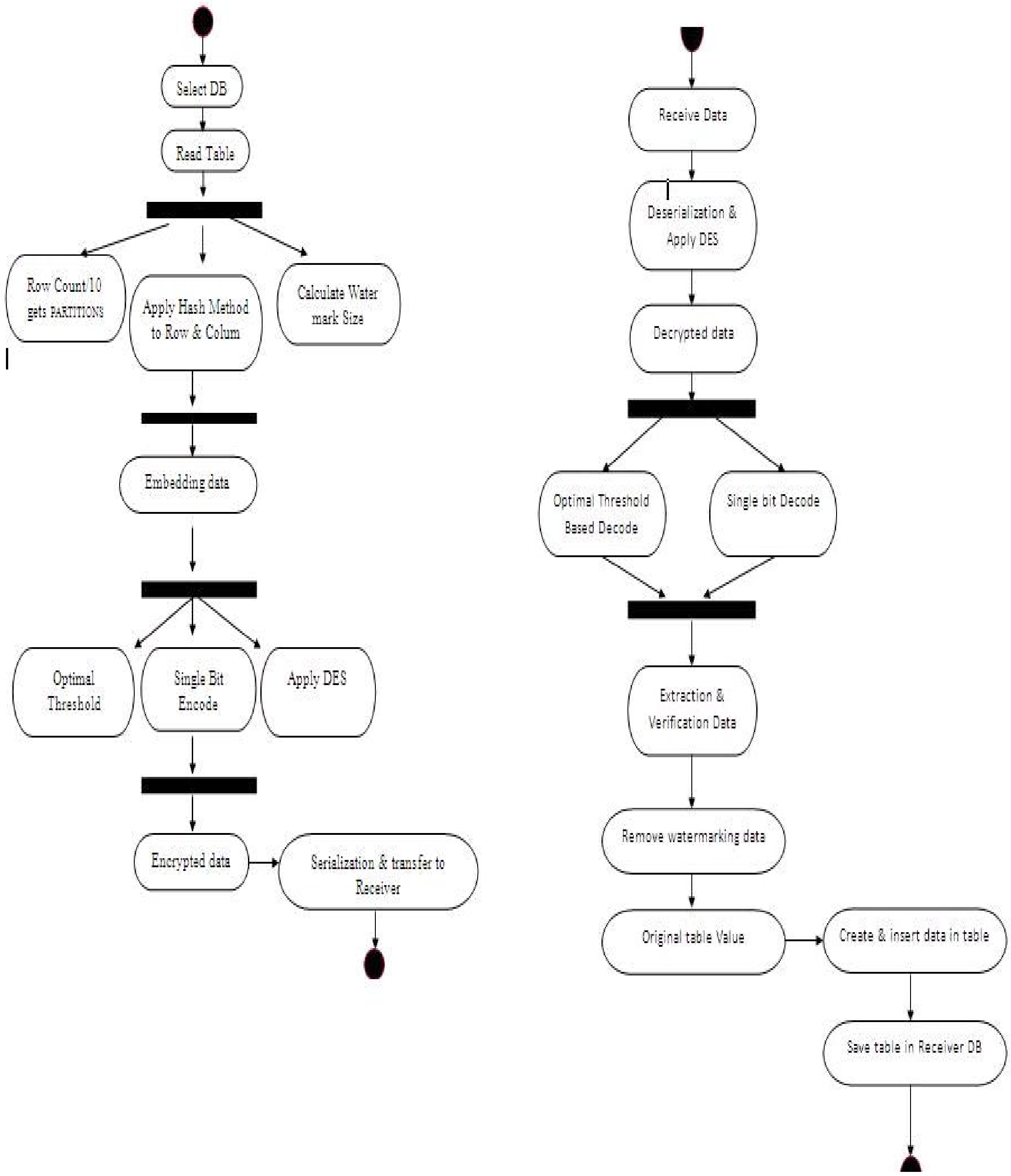

Select DB
Read Table
Row Count/10 gets PARTITIONS
Apply Hash Method to Row & Colum
Calculate Water mark Size
Embedding data
Optimal Threshold
Single Bit Encode
Apply DES
Encrypted data
Serialization & transfer to Receiver
Receive Data
Deserialization & Apply DES
Decrypted data
Optimal Threshold Based Decode
Single bit Decode
Extraction & Verification Data
Remove watermarking data
Original table Value
Create & insert data in table
Save table in Receiver DB

FIGURE: 2 & 3 Entity Relationship Diagram-Sender Activity and Receiver Activity

5. SYSTEM DESIGN

5.1 INPUT DESIGN

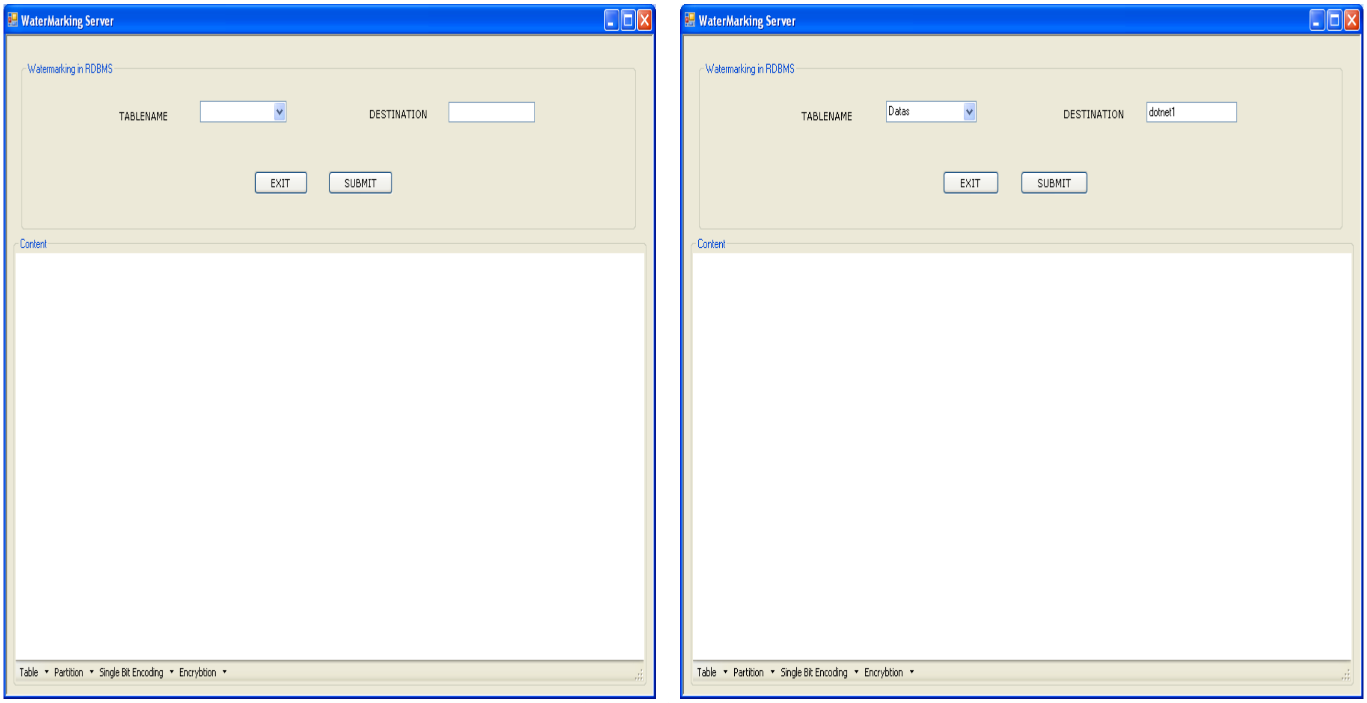
Input Design is the process of converting user oriented inputs to a computer based format. The quality of the system input determines the quality of the system output. Input design determines the format and validation criteria for data entering to the system. Input design is a part of overall system design, which requires very careful attention. If the data going into the system is incorrect then the processing and output will magnifies these errors. The analysis phase should consider the impact of the inputs on the system as a whole and on the other systems. In this project, the inputs are designed in such a way that occurrence of errors are minimized to its maximum since only authorized user are administrator can able to access this tool. The input is given by the administrator are checked at the entry form itself. So there is no chance of unauthorized accessing of the tool. Any abnormally found in the inputs are checked and handled effectively. Input design features can ensure the reliability of a system and produce results from accurate data or they can result in the production of erroneous information.

5.2 OUTPUT DESIGN

Computer output is the most important and direct source of information to the users. Designing the output should proceed in an organized, well thought out manner. The right output must be developed while ensuring that each output element is designed so that people will find easy to use the system. When analysts design the output, they identify the specific output that is needed to meet the information requirements. The success and failure of the system depends on the output, through a system looks attractive and user friendly, the output it produces decides upon the usage of the system. The outputs generated by the system are checked for its consistency, and output is provided simple so that user can handle them with ease. For many end users, output is the main reason for developing the system and the basis on which they will evaluate the usefulness of the application.

6. IMPLEMENTATION

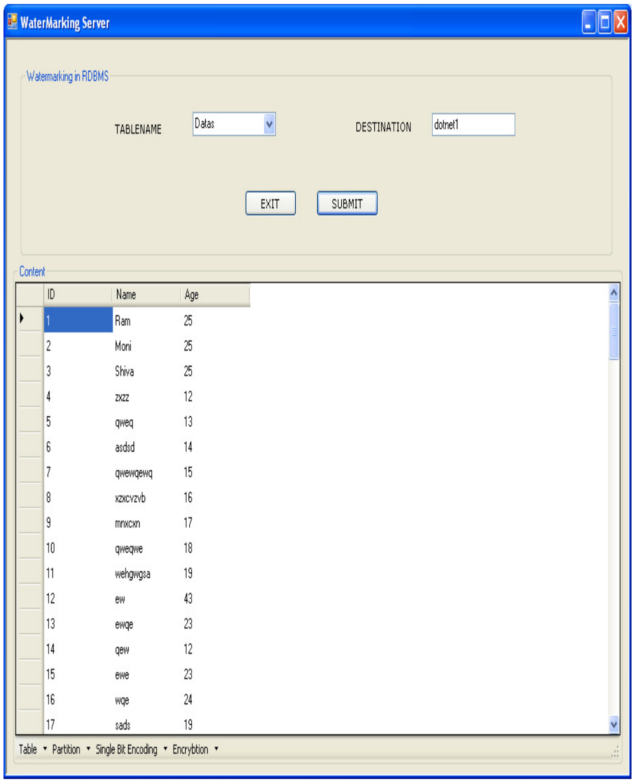
Once the system has been designed, the next step is to convert the designed one in to actual code, so as to satisfy the user requirements as excepted. If the system is approved to be error free it can be implemented.



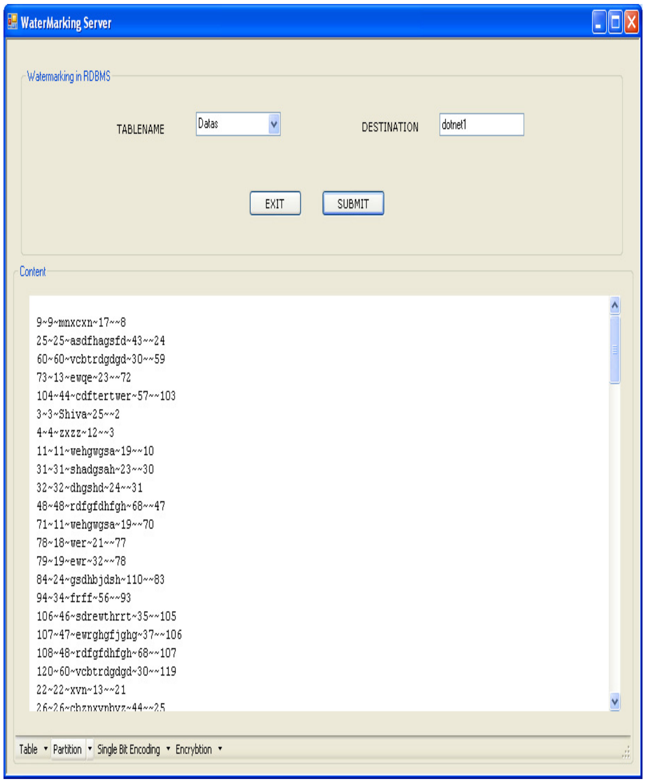
TABLENAME Datas

DESTINATION dotnet1

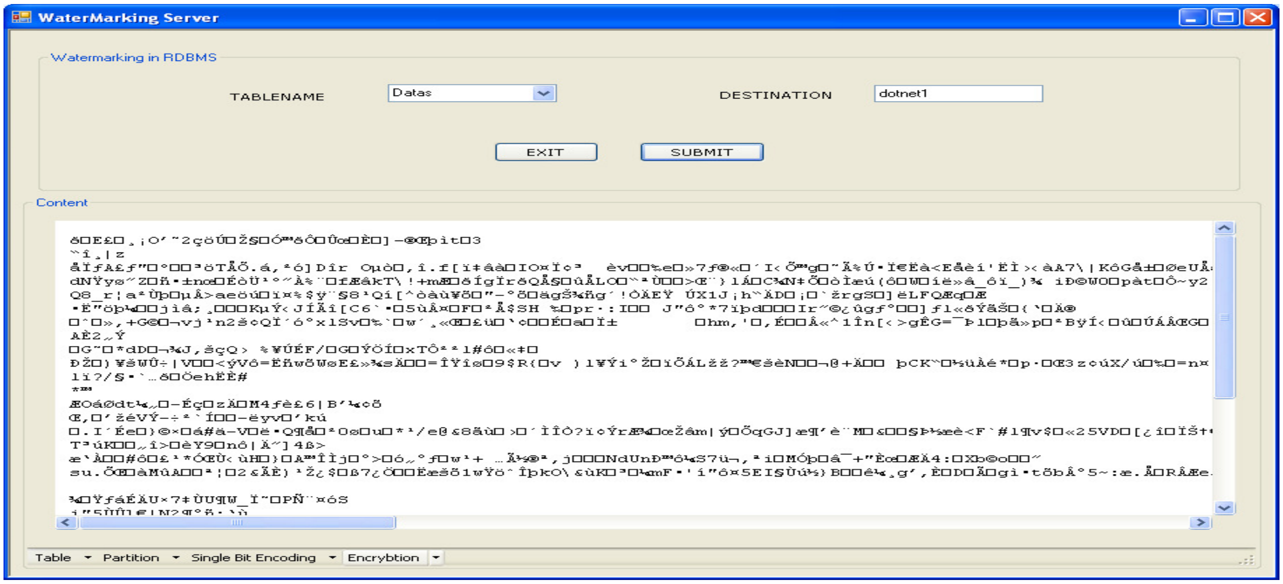
Screen Shots



ID	Name	Age
1	Ram	25
2	Moni	25
3	Shiva	25
4	zxzz	12
5	qweq	13
6	asdsd	14
7	qwewqewq	15
8	xzxcvzvb	16
9	mnxcxn	17
10	qweqwe	18
11	wehgwgsa	19
12	ew	43
13	ewqe	23
14	qew	12
15	ewe	23
16	wqe	24
17	sads	19



```
9~9~mnxcxn~17~~8
25~25~asdfhagsfd~43~~24
60~60~vcbtrdgdgd~30~~59
73~13~ewqe~23~~72
104~44~cdftertwer~57~~103
3~3~Shiva~25~~2
4~4~zxzz~12~~3
11~11~wehgwgsa~19~~10
31~31~shadgsah~23~~30
32~32~dhgshd~24~~31
48~48~rdfgfdhfgh~68~~47
71~11~wehgwgsa~19~~70
78~18~wer~21~~77
79~19~ewr~32~~78
84~24~gsdhbjdsh~110~~83
94~34~frff~56~~93
106~46~sdrewthrrt~35~~105
107~47~ewrghgfjghg~37~~106
108~48~rdfgfdhfgh~68~~107
120~60~vcbtrdgdgd~30~~119
22~22~xvn~13~~21
26~26~chznxvnbvz~44~~25
```



Screen Shots

7. CONCLUSION AND FUTURE ENHANCEMENT

The watermarking problem was formulated as a constrained optimization problem that maximizes or minimizes a hiding function based on the bit to be embedded. GA and PS techniques were employed to solve the proposed optimization problem and to handle the constraints. Furthermore, the data partitioning technique that does not depend on special marker tuples to locate the partitions and proved its resilience to watermark synchronization errors. Development of an efficient threshold-based technique for watermark detection that is based on an optimal threshold that minimizes the probability of decoding error. The watermark resilience was improved by the repeated embedding of the watermark and using majority voting technique in the watermark decoding phase. Moreover, the watermark resilience was improved by using multiple attributes. Media files can be transferred with secure and less packet loss. The ARMS system architecture with a focus on the extensions to the ISMA security standard to enable adaptive streaming of encrypted MPEG-4 content. Investigation on various optimizations in the coding and streaming to improve the bandwidth utilization while minimizing the distortion experienced by the clients in wired and wireless networks is going. Advances in compression and an increase in affordable bandwidth will allow for the streaming of higher resolution video and crisper audio. Developing better speech to text software and more adaptive technologies in streaming will offer greater accessibility.

8. REFERENCES

[1] R. Agrawal and J. Kiernan, "Watermarking Relational Databases," Proc. 28th Int'l Conf. Very Large Data Bases, 2002.

[2] M. Atallah and S. Lonardi, "Authentication of LZ-77 Compressed Data," Proc. ACM Symp. Applied Computing, 2003.

[3] M. Atallah, V. Raskin, C. Hempelman, M. Karahan, R. Sion, K. Triezenberg, and U. Topkara, "Natural Language Watermarking and Tamperproofing," Proc. Fifth Int'l Information Hiding Workshop, 2002.

[4] G. Box, "Evolutionary Operation: A Method for Increasing Industrial Productivity," Applied Statistics, vol. 6, no. 2, pp. 81- 101, 1957.

[5] E. Chong and S. Z_ ak, An Introduction to Optimization. John Wiley & Sons, 2001.

[6] D. Coley, "Introduction to Genetic Algorithms for Scientists and Engineers," World Scientific, 1999.

[7] C. Collberg and C. Thomborson, "Software Watermarking: Models and Dynamic Embeddings," Proc. 26th ACM SIGPLANSIGACT Symp. Principles of Programming Languages, Jan. 1999.

[8] I. Cox, J. Bloom, and M. Miller, Digital Watermarking. Morgan Kaufmann, 2001.

[9] E. Dolan, R. Lewis, and V. Torczon, "On the Local Convergence of Pattern Search," SIAM J. Optimization, vol. 14, no. 2, pp. 567-583, 2003.

[10] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," Proc. IEEE, vol. 87, no. 7, pp. 1079- 1107, July 1999.

[11] Darshana Mistry, "Comparison of Digital Water Marking methods," International Journal on Computer Science and Engineering - Vol. 02, No. 09, 2010, 2905-2909

[12] Dolley Shukla and Manisha Sharma, "WATERMARKING SCHEMES FOR COPY PROTECTION: A SURVEY," International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.1, February 2012

[13] K.Ganesan and Tarun Kumar Guptha, "Multiple Binary Images Watermarking in Spatial and Frequency Domains," Signal & Image Processing: An International Journal (SIPIJ) Vol.1, No.2, December 2010

Automated Data Integration, Cleaning and Analysis Using Data Mining and SPSS Tool For Technical School in Malaysia

Tajul Rosli Razak *tajulrosli@perlis.uitm.edu.my*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(Perlis)
02600 Arau, Perlis, Malaysia.

Abdul Hapes Mohammed *hapes232@perlis.uitm.edu.my*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(Perlis)
02600 Arau, Perlis, Malaysia.

Noorfaizalfarid Hj Mohd Noor *noorfaizal455@perlis.uitm.edu.my*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(Perlis)
02600 Arau, Perlis, Malaysia.

Muhamad Arif Hashim *muhamadarif487@perlis.uitm.edu.my*
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA(Perlis)
02600 Arau, Perlis, Malaysia.

Abstract

Students' performance plays major role in determining the quality of our education system. Sijil Pelajaran Malaysia (SPM) is a public examination compulsory to be taken by Form 5 students in Malaysia. The performance gap is not only a school and classroom issue but also a national issue that must be addressed properly. This study aims to integrate, clean and analysis through automated data mining techniques. Using data mining techniques is one of the processes of transferring raw data from current educational system to meaningful information that can be used to help the school community to make a right decision to achieve much better results. This proved DM provides means to assist both educators and students, and improve the quality of education. The result and findings in the study show that automated system will give the same result compare with manual system of integration and analysis and also could be used by the management to make faster and more efficient decision in order to map or plan efficient teaching approach for students in the future.

Keywords: Data Integration, Data Cleaning, Data Analysis, Decision Support

1. INTRODUCTION

Examinations serve many purposes, which are to make assessment on the effectiveness of our education process, and subsequently facilitate improvement on the process. Examinations also serve the function of differentiations among students so that different groups of student with unique level learning ability can be grouped together for differentiated education.

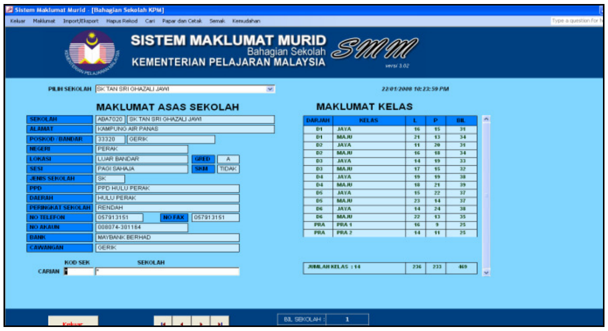
Education is viewed as a critical factor in contributing to the long-term economic well-being of the country. Therefore, government realizes that the importance of maximizing the potential of each individual student, as well as the education system. In Malaysia, students generally are eligible in pursuing their higher education learning after finishing the secondary schools. From there, students have choices on how to pursuing their studies, either to join higher learning institutions such as polytechnic, public universities, private universities, community college or university college to enroll the diploma or certificate level. On the other hand, students also may undertake the upper secondary program for two years in school and sit for Sijil Pelajaran Malaysia (SPM).

Students' performance plays major role in determining the quality of our education system. Sijil Pelajaran Malaysia (SPM) is a public examination compulsory to be taken by Form 5 students in Malaysia. The performance gap is not only a school and classroom issue but also a national issue that must be addressed properly [1]. As young generation, their performance in school is important to determine which schools that these children are streamed to further their studies either to daily school, boarding school, semi boarding school or religious school. In effect, this will influence their career path in the future. Factors such as gender [2,3], attendance [1,4], co-curricular activities [5] and family background [6] may influence their performance.

Generally, this study aims to integrate, clean and used that data for automated analysis through data mining techniques. Using data mining techniques is one of the processes of transferring raw data from current educational system to meaningful information that can be used to help the school community to make a right decision to achieve much better results. This proved DM provides means to assist both educators and students, and improve the quality of education. Unfortunately, using the traditional method not only increase the teaching load of the teachers, but also gives unnecessary burdens to students [7]. The result and findings in the study could be used by the management to make faster and more efficient decision in order to map or plan efficient teaching approach for students in the future.

2. PROBLEM STATEMENTS

Currently, most schools in Malaysia use Sistem Maklumat Murid (SMM) to collect their information related to family background, income and others as shown in Fig 1. The system gathered information including their names, birth certificate numbers, gender, age, parent's name, parent's job, parent's income, guidance status and sibilings. In order to get detail information about students, it is important to access to 'Borang Maklumat Murid' (BMM).



PILIH SEKOLAH: SK TAN SRI GHAZALI JAWI

22/01/2008 10:23:59 PM

MAKLUMAT ASAS SEKOLAH

SEKOLAH: ABA7020 SK TAN SRI GHAZALI JAWI
ALAMAT: KAMPUNG AIR PANAS
POSKOD / BANDAR: 33320 GERIK
NEGERI: PERAK
LOKASI: LUAR BANDAR GRED: A
SESI: PAGI SAHAJA SKM: TIDAK
JENIS SEKOLAH: SK
PPD: PPD HULU PERAK
DAERAH: HULU PERAK
PERINGKAT SEKOLAH: RENDAH
NO TELEFON: 057913151 NO FAX: 057913151
NO AKAUN: 008074-301164
BANK: MAYBANK BERHAD
CAWANGAN: GERIK

CARIAN KOD SEK: SEKOLAH: *

MAKLUMAT KELAS

DARJAH	KELAS	L	P	BIL
D1	JAYA	16	15	31
D1	MAJU	21	13	34
D2	JAYA	11	20	31
D2	MAJU	16	18	34
D3	JAYA	14	19	33
D3	MAJU	17	15	32
D4	JAYA	19	19	38
D4	MAJU	18	21	39
D5	JAYA	15	22	37
D5	MAJU	23	14	37
D6	JAYA	14	24	38
D6	MAJU	22	13	35
PRA	PRA 1	16	9	25
PRA	PRA 2	14	11	25
JUMLAH KELAS : 14		236	233	469

FIGURE 1: Interface of SMM

Monthly examination marks are normally keyed-in by respective teacher and stored in Microsoft Excel. Analyses such as crosstabulation and prediction model development have not been explored since data has been key-in independently. School in general is rich with data which is beneficial if the later could be used to help teacher and management understand more about their student background based on their performance. Due to lack of effort in integrating table or database between SMM and student result, this study attempts to uncover the hidden information within SMM data and student result. This study also takes initiative to assist teacher upload their data file in the server for integration and analysis purposes will be done automatically by using automated web based with data mining facilities.

3. OBJECTIVE OF STUDY

Generally, the main objectives of this study are to perform integration, cleaning and automated analysis on school data management by using data mining approaches. The specific objectives are listed as:

i. To integrate databases from different sources.
ii. To preprocess data prior to mining process.
iii. To design and implement the prototype of automated data integration.
iv. To evaluate integrated data mining using data mining methods.

4. METHODOLOGY

The process flow of the study is illustrated as shown in Fig. 2 that consists of stage Integration, Extract, Cleaning.

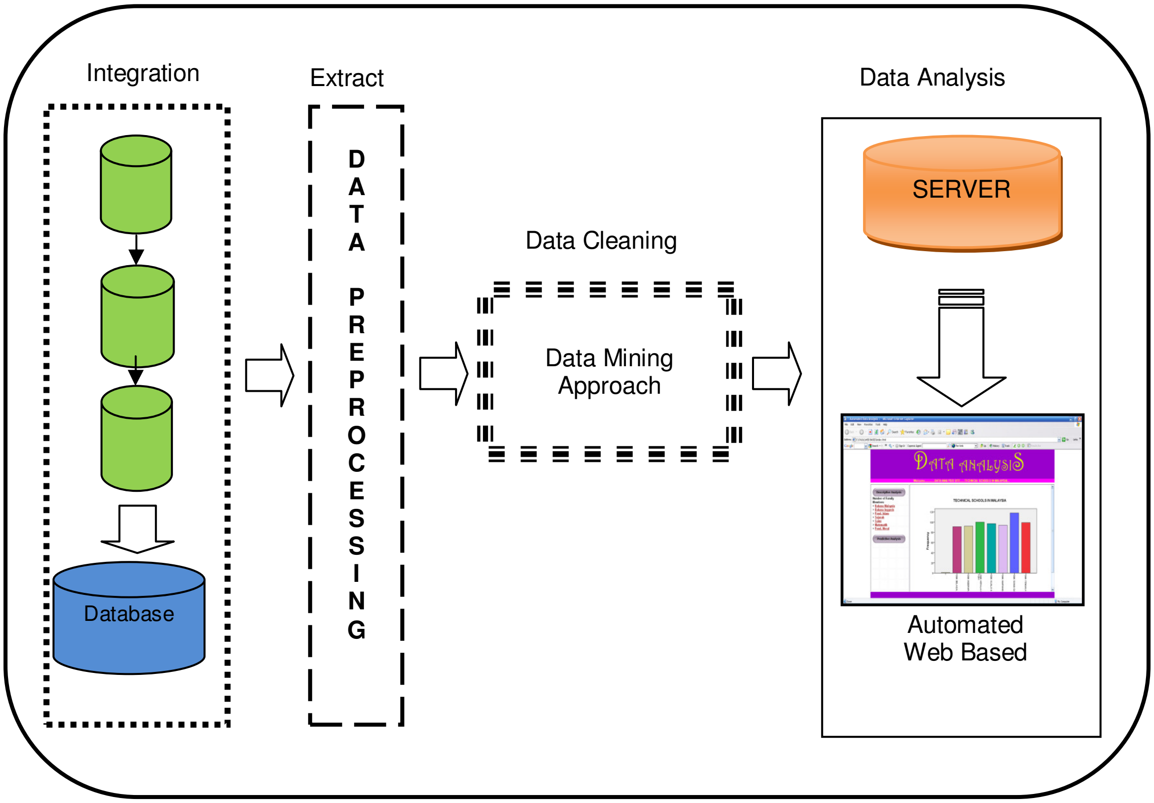


FIGURE 2: Process Flow of the study

4.1 Data Integration

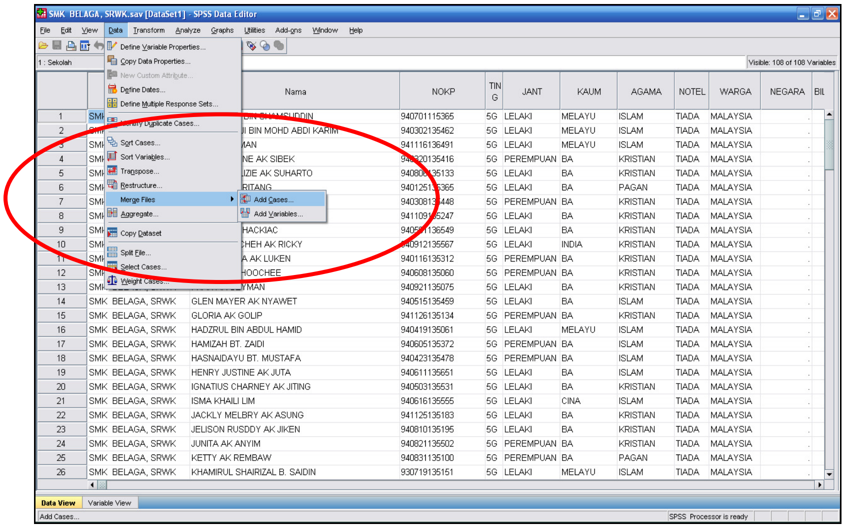
The initial phase is concerned with collection of data in Microsoft Excel format that integrate seven technical schools in Malaysia. These technical schools include SMK BELAGA SARAWAK, SMK INDERAPURA PAHANG, SMK KEPALA BATAS KEDAH, SMK KUALA KETIL KEDAH, SMK MARANG TERENGGANU, SMK SAMA GAGAH PULAU PINANG, and SMK TENGKU IDRIS SELANGOR. The data will be collected manually and was assist by Jabatan Pelajaran Negeri of each state. These technical schools were show in table 1 below. To analyze the results, descriptive statistics such as frequencies, cross tabulation, and charts were used to describe the output.

TECHNICAL SCHOOLS	STATE
SMK BELAGA	SARAWAK
SMK INDERAPURA	PAHANG
SMK KEPALA BATAS	KEDAH
SMK KUALA KETIL	KEDAH
SMK MARANG	TERENGGANU
SMK SAMA GAGAH	PULAU PINANG
SMK TENGKU IDRIS	SELANGOR

Table 1: Technical Schools that was selected to be samples

4.2 Instrument

To make integration of that data, software SPSS 16.0 has been used to combine seven technical schools. The snapshot of the integration process is shown in Fig. 3.



1 : Sekolah

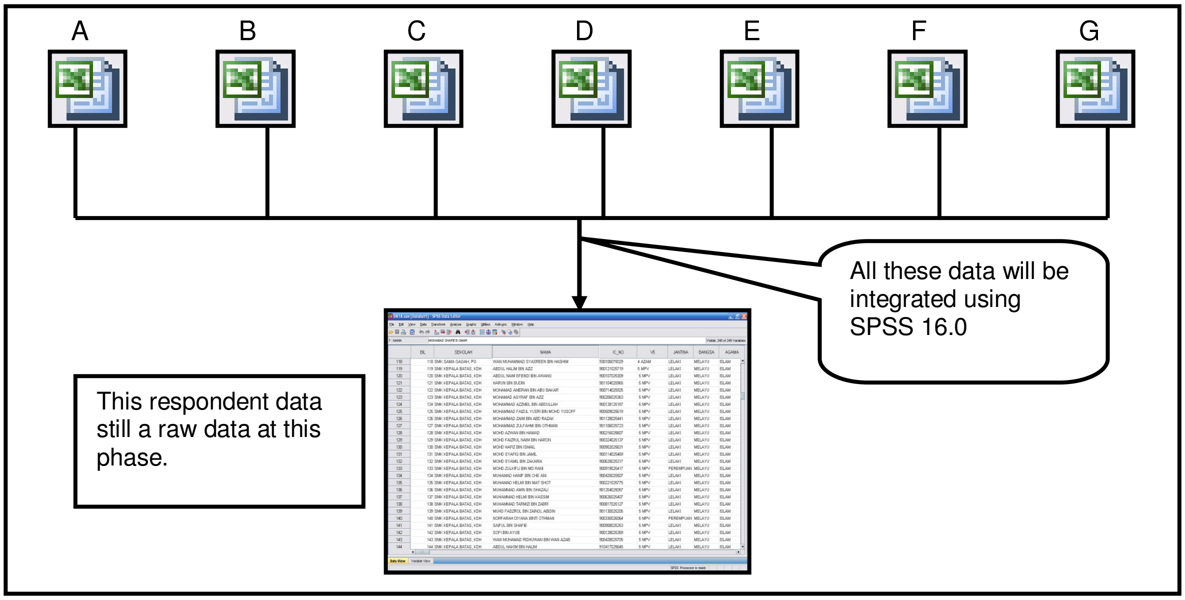
Visible: 108 of 108 Variables

		Nama	NOKP	TING	JANT	KAUM	AGAMA	NOTEL	WARGA	NEGARA	BIL
1	SMK	BIN SHAMSUDDIN	940701115365	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA		.
2	SMK	I BIN MOHD ABDI KARIM	940302135462	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA		.
3	SMK	MAN	941116136491	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA		.
4	SMK	NE AK SIBEK	940320135416	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
5	SMK	IZIE AK SUHARTO	940808135133	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
6	SMK	RITANG	940125135365	5G	LELAKI	BA	PAGAN	TIADA	MALAYSIA		.
7	SMK		940308135448	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
8	SMK		941109135247	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
9	SMK	HACKIAC	940501136549	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
10	SMK	CHEH AK RICKY	940912135567	5G	LELAKI	INDIA	KRISTIAN	TIADA	MALAYSIA		.
11	SMK	A AK LUKEN	940116135312	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
12	SMK	HOOCHEE	940608135060	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
13	SMK	YMAN	940921135075	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
14	SMK BELAGA, SRWK	GLEN MAYER AK NYAWET	940515135459	5G	LELAKI	BA	ISLAM	TIADA	MALAYSIA		.
15	SMK BELAGA, SRWK	GLORIA AK GOLIP	941126135134	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
16	SMK BELAGA, SRWK	HADZRUL BIN ABDUL HAMID	940419135061	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA		.
17	SMK BELAGA, SRWK	HAMIZAH BT. ZAIDI	940605135372	5G	PEREMPUAN	BA	ISLAM	TIADA	MALAYSIA		.
18	SMK BELAGA, SRWK	HASNAIDAYU BT. MUSTAFA	940423135478	5G	PEREMPUAN	BA	ISLAM	TIADA	MALAYSIA		.
19	SMK BELAGA, SRWK	HENRY JUSTINE AK JUTA	940611135651	5G	LELAKI	BA	ISLAM	TIADA	MALAYSIA		.
20	SMK BELAGA, SRWK	IGNATIUS CHARNEY AK JITING	940503135531	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
21	SMK BELAGA, SRWK	ISMA KHAILI LIM	940616135555	5G	LELAKI	CINA	ISLAM	TIADA	MALAYSIA		.
22	SMK BELAGA, SRWK	JACKLY MELBRY AK ASUNG	941125135183	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
23	SMK BELAGA, SRWK	JELISON RUSDDY AK JIKEN	940810135195	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA		.
24	SMK BELAGA, SRWK	JUNITA AK ANYIM	940821135502	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA		.
25	SMK BELAGA, SRWK	KETTY AK REMBAW	940831135100	5G	PEREMPUAN	BA	PAGAN	TIADA	MALAYSIA		.
26	SMK BELAGA, SRWK	KHAMIRUL SHAIRIZAL B. SAIDIN	930719135151	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA		.

FIGURE 3 : Snapshot of the integration process

4.3 Respondents Data

There are seven technical schools in Malaysia that has been collected to represent as respondent data for this study which is SMK BELAGA SARAWAK (A), SMK INDERAPURA PAHANG (B), SMK KEPALA BATAS KEDAH (C), SMK KUALA KETIL KEDAH (D), SMK MARANG TERENGGANU (E), SMK SAMA GAGAH PULAU PINANG (F), and SMK TENGKU IDRIS SELANGOR (G) and all come in one format that is Microsoft Excel. These data must be integrated together by using SPSS 16.0 software and the flow of process was show in Fig 4 below.



All these data will be integrated using SPSS 16.0

This respondent data still a raw data at this phase.

DATA.sav [DataSet1] - SPSS Data Editor

7 : NAMA MOHAMAD SHAFIE B OMAR

Visible: 245 of 245 Variables

	BIL	SEKOLAH	NAMA	IC_NO	V5	JANTINA	BANGSA	AGAMA
118	118	SMK SAMA GAGAH, PG	WAN MUHAMMAD SYASREEN BIN HASHIM	930106075029	4 AZAM	LELAKI	MELAYU	ISLAM
119	119	SMK KEPALA BATAS, KDH	ABDUL HALIM BIN AZIZ	900121025719	5 MPV	LELAKI	MELAYU	ISLAM
120	120	SMK KEPALA BATAS, KDH	ABDUL NAIM EFENDI BIN AWANG	900107025309	5 MPV	LELAKI	MELAYU	ISLAM
121	121	SMK KEPALA BATAS, KDH	HARUN BIN BUDIN	901104025965	5 MPV	LELAKI	MELAYU	ISLAM
122	122	SMK KEPALA BATAS, KDH	MOHAMAD AMERAN BIN ABU BAKAR	900714025925	5 MPV	LELAKI	MELAYU	ISLAM
123	123	SMK KEPALA BATAS, KDH	MOHAMAD ASYRAF BIN AZIZ	900206025363	5 MPV	LELAKI	MELAYU	ISLAM
124	124	SMK KEPALA BATAS, KDH	MOHAMMAD AZZMEL BIN ABDULLAH	900128125157	5 MPV	LELAKI	MELAYU	ISLAM
125	125	SMK KEPALA BATAS, KDH	MOHAMMAD FAIZUL YUSRI BIN MOHD YUSOFF	900509025619	5 MPV	LELAKI	MELAYU	ISLAM
126	126	SMK KEPALA BATAS, KDH	MOHAMMAD ZAIM BIN ABD RAZAK	901128025441	5 MPV	LELAKI	MELAYU	ISLAM
127	127	SMK KEPALA BATAS, KDH	MOHAMMAD ZULFAHMI BIN OTHMAN	901106025723	5 MPV	LELAKI	MELAYU	ISLAM
128	128	SMK KEPALA BATAS, KDH	MOHD AZWAN BIN HAMAD	900216025607	5 MPV	LELAKI	MELAYU	ISLAM
129	129	SMK KEPALA BATAS, KDH	MOHD FAIZRUL NAIM BIN HARON	900324025137	5 MPV	LELAKI	MELAYU	ISLAM
130	130	SMK KEPALA BATAS, KDH	MOHD HAFIZ BIN ISMAIL	900902025631	5 MPV	LELAKI	MELAYU	ISLAM
131	131	SMK KEPALA BATAS, KDH	MOHD SYAFIQ BIN JAMIL	900114025469	5 MPV	LELAKI	MELAYU	ISLAM
132	132	SMK KEPALA BATAS, KDH	MOHD SYAMIL BIN ZAKARIA	900629025317	5 MPV	LELAKI	MELAYU	ISLAM
133	133	SMK KEPALA BATAS, KDH	MOHD ZULKIFLI BIN MD RANI	900919025417	5 MPV	PEREMPUAN	MELAYU	ISLAM
134	134	SMK KEPALA BATAS, KDH	MUHAMAD HANIF BIN CHE ANI	900425025507	5 MPV	LELAKI	MELAYU	ISLAM
135	135	SMK KEPALA BATAS, KDH	MUHAMAD HELMI BIN MAT SHOT	900221025775	5 MPV	LELAKI	MELAYU	ISLAM
136	136	SMK KEPALA BATAS, KDH	MUHAMMAD AMIN BIN GHAZALI	901204025057	5 MPV	LELAKI	MELAYU	ISLAM
137	137	SMK KEPALA BATAS, KDH	MUHAMMAD HELMI BIN KASSIM	900626025407	5 MPV	LELAKI	MELAYU	ISLAM
138	138	SMK KEPALA BATAS, KDH	MUHAMMAD TARMIZI BIN ZABRI	900817025127	5 MPV	LELAKI	MELAYU	ISLAM
139	139	SMK KEPALA BATAS, KDH	MUHD FADZROL BIN ZAINOL ABIDIN	901130025205	5 MPV	LELAKI	MELAYU	ISLAM
140	140	SMK KEPALA BATAS, KDH	NORFARAH DIYANA BINTI OTHMAN	900330026064	5 MPV	PEREMPUAN	MELAYU	ISLAM
141	141	SMK KEPALA BATAS, KDH	SAIFUL BIN SHAFIE	900908025253	5 MPV	LELAKI	MELAYU	ISLAM
142	142	SMK KEPALA BATAS, KDH	SOFI BIN AYUB	900128025359	5 MPV	LELAKI	MELAYU	ISLAM
143	143	SMK KEPALA BATAS, KDH	WAN MUHAMAD RIDHUWAN BIN WAN AZAB	900428025705	5 MPV	LELAKI	MELAYU	ISLAM
144	144	SMK KEPALA BATAS, KDH	ABDUL HAKIM BIN HALIM	910417025645	5 MPV	LELAKI	MELAYU	ISLAM

SPSS Processor is ready

FIGURE 4: Flow of Integration in SPSS 16.0 Software

After the integration process, this respondent data are included all technical schools together and 691 respondent have successfully obtained. Before this samples will be process to the next phase which is data cleaning process, this data of all technical schools are show in Table 2.

Technical Schools	Frequency	%
SMK BELAGA, SRWK	91	13.2
SMK INDERAPURA, PHG	92	13.3
SMK KEPALA BATAS, KDH	100	14.5
SMK KUALA KETIL, KDH	97	14.0
SMK MARANG, TRG	94	13.6
SMK SAMA GAGAH, PG	118	17.1
SMK TENGKU IDRIS, SLGR	99	14.3
Total	**691**	**100.0**

TABLE 2: Respondent Data

4.4 Target

According to Shmueli [8], the target dependent variable can be denoted as one of the attributes being used to predict in supervised learning. In this study, SPM attribute's grade is used as the target. SPM attribute's grade was categorized into 5 groups as shown in Table 3.

SPM Grade	Class
1A ~ 2A	1
3B ~ 4B	2
5C ~ 6C	3
7D ~ 8E	4
9G	5

TABLE 3: SPM Grade categorized

4.5 Data Cleaning

The second phase which is extract and cleaning will be run together and will used data mining approach. Database that is stored in first phase maybe will have high probability of 'dirty data'. Data will be extracted from database and its need to do data preprocesses and then does cleaning. Data cleaning approach should satisfy several requirements. This process was carried out using SPSS version 16.0. Data selection was carried out in this phase and its purpose is to ensure that predicting model can produce more accurate results. Table 4 shows target and various attributes selected before converted to the purposes analysis.

Type	Input Variable	Domain
Target	SPM subjects	1A,2A,3B,4B,5C,6C,7D,8E,9G
Attributes	Number of family members	Numeric
Attributes	Number of family member still learning	Numeric
Attributes	Number of family member receive SPBT	Numeric
Attributes	Family Income	RM 500, RM 1200, RM 1900
Attributes	SPBT	YA, TIDAK

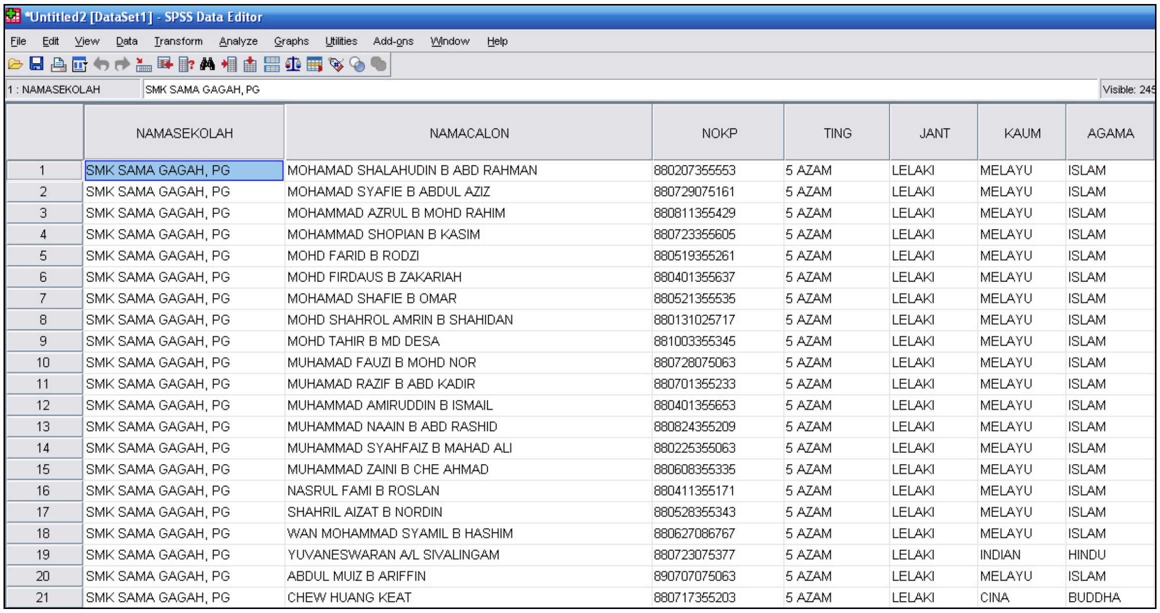
TABLE 4: The selected attributes before converted to numeric

Table 5 shows target and various attributes selected after converted for this analysis process. Then, the raw dataset will be changing into numerical forms for experiment sake.

Type	Input Variable	Domain
Target	SPM subjects	1, 2, 3, 4, 5
Attributes	Number of family members	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
Attributes	Number of family member still learning	0, 1, 2, 3, 4, 5, 6, 7
Attributes	Number of family member receive SPBT	0, 1, 2, 3, 4, 5, 6 7
Attributes	Total Income	1, 2, 3, 4, 5, 6
Attributes	SPBT	1, 2

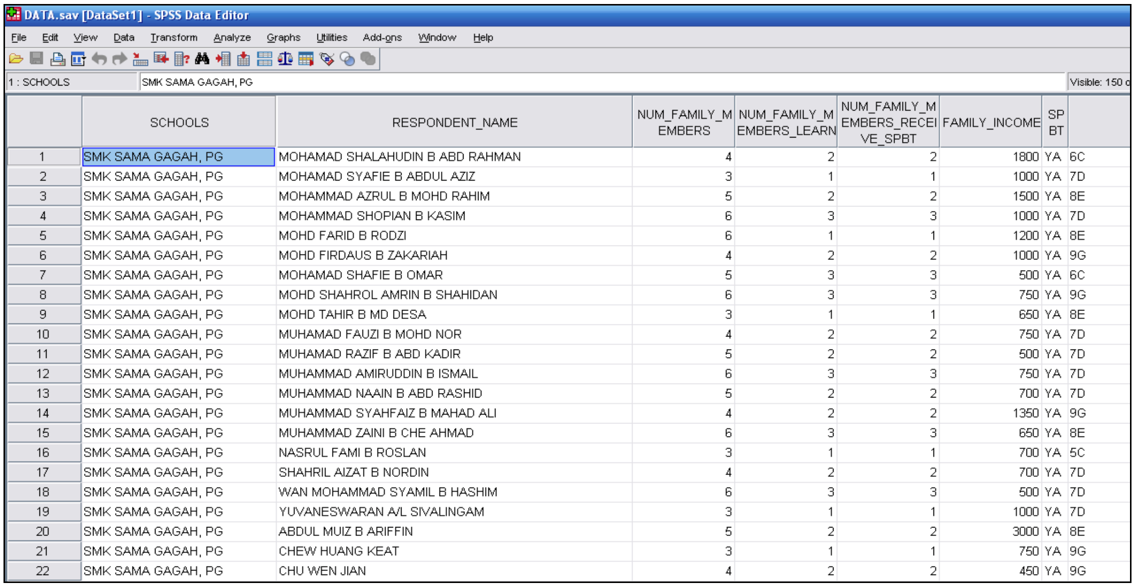
TABLE 5: The selected attributes after converted to numeric

Meanwhile, the process of data transformation was done to ensure the data formats are in appropriate form that can using by SPSS 16.0 tool. (See Fig. 5 to Fig. 8)



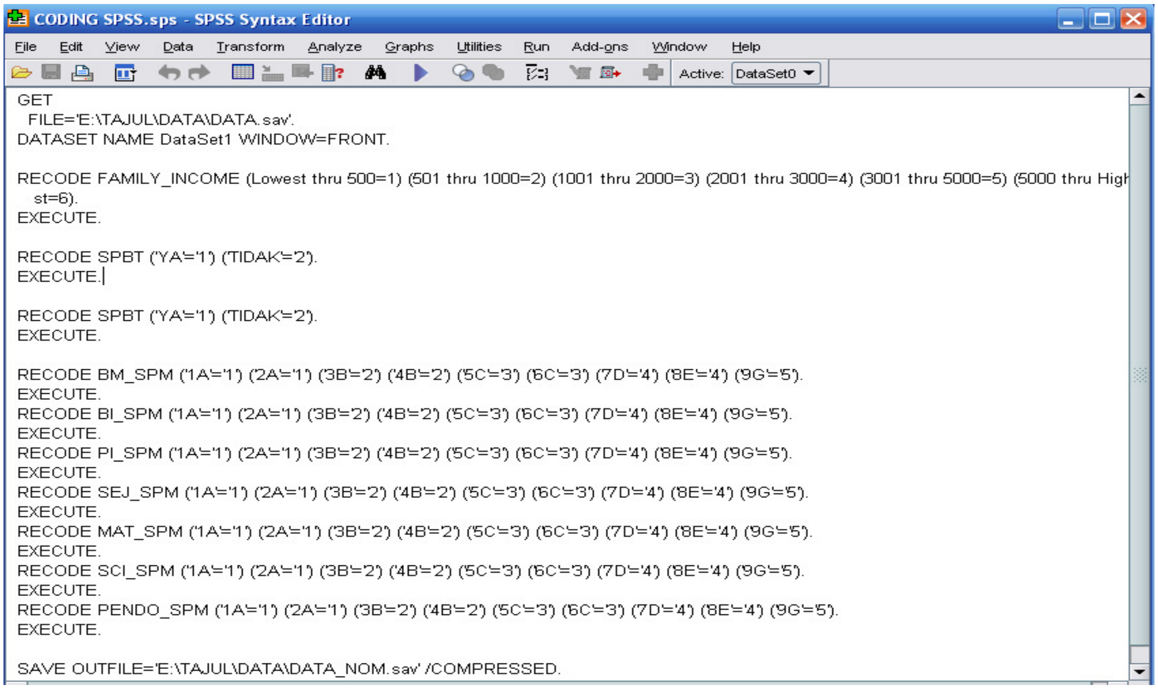
	NAMASEKOLAH	NAMACALON	NOKP	TING	JANT	KAUM	AGAMA
1	SMK SAMA GAGAH, PG	MOHAMAD SHALAHUDIN B ABD RAHMAN	880207355553	5 AZAM	LELAKI	MELAYU	ISLAM
2	SMK SAMA GAGAH, PG	MOHAMAD SYAFIE B ABDUL AZIZ	880729075161	5 AZAM	LELAKI	MELAYU	ISLAM
3	SMK SAMA GAGAH, PG	MOHAMMAD AZRUL B MOHD RAHIM	880811355429	5 AZAM	LELAKI	MELAYU	ISLAM
4	SMK SAMA GAGAH, PG	MOHAMMAD SHOPIAN B KASIM	880723355605	5 AZAM	LELAKI	MELAYU	ISLAM
5	SMK SAMA GAGAH, PG	MOHD FARID B RODZI	880519355261	5 AZAM	LELAKI	MELAYU	ISLAM
6	SMK SAMA GAGAH, PG	MOHD FIRDAUS B ZAKARIAH	880401355637	5 AZAM	LELAKI	MELAYU	ISLAM
7	SMK SAMA GAGAH, PG	MOHAMAD SHAFIE B OMAR	880521355535	5 AZAM	LELAKI	MELAYU	ISLAM
8	SMK SAMA GAGAH, PG	MOHD SHAHROL AMRIN B SHAHIDAN	880131025717	5 AZAM	LELAKI	MELAYU	ISLAM
9	SMK SAMA GAGAH, PG	MOHD TAHIR B MD DESA	881003355345	5 AZAM	LELAKI	MELAYU	ISLAM
10	SMK SAMA GAGAH, PG	MUHAMAD FAUZI B MOHD NOR	880728075063	5 AZAM	LELAKI	MELAYU	ISLAM
11	SMK SAMA GAGAH, PG	MUHAMAD RAZIF B ABD KADIR	880701355233	5 AZAM	LELAKI	MELAYU	ISLAM
12	SMK SAMA GAGAH, PG	MUHAMMAD AMIRUDDIN B ISMAIL	880401355653	5 AZAM	LELAKI	MELAYU	ISLAM
13	SMK SAMA GAGAH, PG	MUHAMMAD NAAIN B ABD RASHID	880824355209	5 AZAM	LELAKI	MELAYU	ISLAM
14	SMK SAMA GAGAH, PG	MUHAMMAD SYAHFAIZ B MAHAD ALI	880225355063	5 AZAM	LELAKI	MELAYU	ISLAM
15	SMK SAMA GAGAH, PG	MUHAMMAD ZAINI B CHE AHMAD	880608355335	5 AZAM	LELAKI	MELAYU	ISLAM
16	SMK SAMA GAGAH, PG	NASRUL FAMI B ROSLAN	880411355171	5 AZAM	LELAKI	MELAYU	ISLAM
17	SMK SAMA GAGAH, PG	SHAHRIL AIZAT B NORDIN	880528355343	5 AZAM	LELAKI	MELAYU	ISLAM
18	SMK SAMA GAGAH, PG	WAN MOHAMMAD SYAMIL B HASHIM	880627086767	5 AZAM	LELAKI	MELAYU	ISLAM
19	SMK SAMA GAGAH, PG	YUVANESWARAN A/L SIVALINGAM	880723075377	5 AZAM	LELAKI	INDIAN	HINDU
20	SMK SAMA GAGAH, PG	ABDUL MUIZ B ARIFFIN	890707075063	5 AZAM	LELAKI	MELAYU	ISLAM
21	SMK SAMA GAGAH, PG	CHEW HUANG KEAT	880717355203	5 AZAM	LELAKI	CINA	BUDDHA

FIGURE 5: Sample of raw dataset before select attributes and converted using SPSS 16.0



	SCHOOLS	RESPONDENT_NAME	NUM_FAMILY_MEMBERS	NUM_FAMILY_MEMBERS_LEARN	NUM_FAMILY_MEMBERS_RECEIVE_SPBT	FAMILY_INCOME	SPBT	
1	SMK SAMA GAGAH, PG	MOHAMAD SHALAHUDIN B ABD RAHMAN	4	2	2	1800	YA	6C
2	SMK SAMA GAGAH, PG	MOHAMAD SYAFIE B ABDUL AZIZ	3	1	1	1000	YA	7D
3	SMK SAMA GAGAH, PG	MOHAMMAD AZRUL B MOHD RAHIM	5	2	2	1500	YA	8E
4	SMK SAMA GAGAH, PG	MOHAMMAD SHOPIAN B KASIM	6	3	3	1000	YA	7D
5	SMK SAMA GAGAH, PG	MOHD FARID B RODZI	6	1	1	1200	YA	8E
6	SMK SAMA GAGAH, PG	MOHD FIRDAUS B ZAKARIAH	4	2	2	1000	YA	9G
7	SMK SAMA GAGAH, PG	MOHAMAD SHAFIE B OMAR	5	3	3	500	YA	6C
8	SMK SAMA GAGAH, PG	MOHD SHAHROL AMRIN B SHAHIDAN	6	3	3	750	YA	9G
9	SMK SAMA GAGAH, PG	MOHD TAHIR B MD DESA	3	1	1	650	YA	8E
10	SMK SAMA GAGAH, PG	MUHAMAD FAUZI B MOHD NOR	4	2	2	750	YA	7D
11	SMK SAMA GAGAH, PG	MUHAMAD RAZIF B ABD KADIR	5	2	2	500	YA	7D
12	SMK SAMA GAGAH, PG	MUHAMMAD AMIRUDDIN B ISMAIL	6	3	3	750	YA	7D
13	SMK SAMA GAGAH, PG	MUHAMMAD NAAIN B ABD RASHID	5	2	2	700	YA	7D
14	SMK SAMA GAGAH, PG	MUHAMMAD SYAHFAIZ B MAHAD ALI	4	2	2	1350	YA	9G
15	SMK SAMA GAGAH, PG	MUHAMMAD ZAINI B CHE AHMAD	6	3	3	650	YA	8E
16	SMK SAMA GAGAH, PG	NASRUL FAMI B ROSLAN	3	1	1	700	YA	5C
17	SMK SAMA GAGAH, PG	SHAHRIL AIZAT B NORDIN	4	2	2	700	YA	7D
18	SMK SAMA GAGAH, PG	WAN MOHAMMAD SYAMIL B HASHIM	6	3	3	500	YA	7D
19	SMK SAMA GAGAH, PG	YUVANESWARAN A/L SIVALINGAM	3	1	1	1000	YA	7D
20	SMK SAMA GAGAH, PG	ABDUL MUIZ B ARIFFIN	5	2	2	3000	YA	8E
21	SMK SAMA GAGAH, PG	CHEW HUANG KEAT	3	1	1	750	YA	9G
22	SMK SAMA GAGAH, PG	CHU WEN JIAN	4	2	2	450	YA	9G

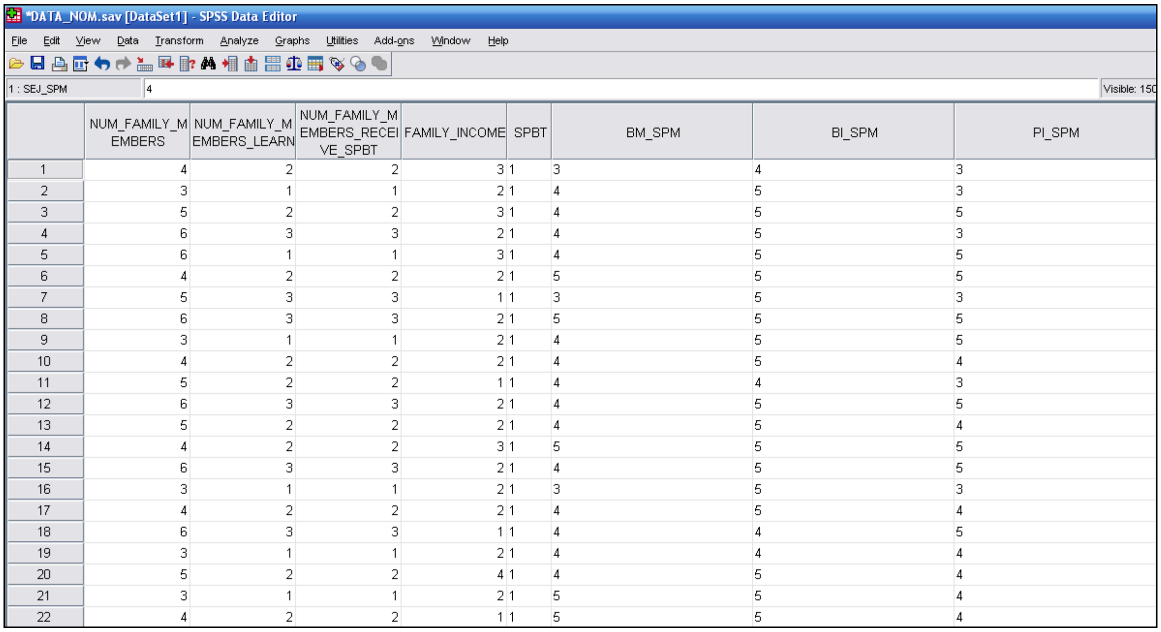
FIGURE 6: The selected attributes



CODING SPSS.sps - SPSS Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Add-ons Window Help
Active: DataSet0

```
GET
  FILE='E:\TAJUL\DATA\DATA.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

RECODE FAMILY_INCOME (Lowest thru 500=1) (501 thru 1000=2) (1001 thru 2000=3) (2001 thru 3000=4) (3001 thru 5000=5) (5000 thru High
  st=6).
EXECUTE.

RECODE SPBT ('YA'='1') ('TIDAK'='2').
EXECUTE.

RECODE SPBT ('YA'='1') ('TIDAK'='2').
EXECUTE.

RECODE BM_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE BI_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE PI_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE SEJ_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE MAT_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE SCI_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.
RECODE PENDO_SPM ('1A'='1') ('2A'='1') ('3B'='2') ('4B'='2') ('5C'='3') ('6C'='3') ('7D'='4') ('8E'='4') ('9G'='5').
EXECUTE.

SAVE OUTFILE='E:\TAJUL\DATA\DATA_NOM.sav' /COMPRESSED.
```

FIGURE 7: Process of data set conversion using Syntax SPSS in SPSS 16.0



*DATA_NOM.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : SEJ_SPM
4
Visible: 150

	NUM_FAMILY_MEMBERS	NUM_FAMILY_MEMBERS_LEARN	NUM_FAMILY_MEMBERS_RECEIVE_SPBT	FAMILY_INCOME	SPBT	BM_SPM	BI_SPM	PI_SPM
1	4	2	2	3	1	3	4	3
2	3	1	1	2	1	4	5	3
3	5	2	2	3	1	4	5	5
4	6	3	3	2	1	4	5	3
5	6	1	1	3	1	4	5	5
6	4	2	2	2	1	5	5	5
7	5	3	3	1	1	3	5	3
8	6	3	3	2	1	5	5	5
9	3	1	1	2	1	4	5	5
10	4	2	2	2	1	4	5	4
11	5	2	2	1	1	4	4	3
12	6	3	3	2	1	4	5	5
13	5	2	2	2	1	4	5	4
14	4	2	2	3	1	5	5	5
15	6	3	3	2	1	4	5	5
16	3	1	1	2	1	3	5	3
17	4	2	2	2	1	4	5	4
18	6	3	3	1	1	4	4	5
19	3	1	1	2	1	4	4	4
20	5	2	2	4	1	4	5	4
21	3	1	1	2	1	5	5	4
22	4	2	2	1	1	5	5	4

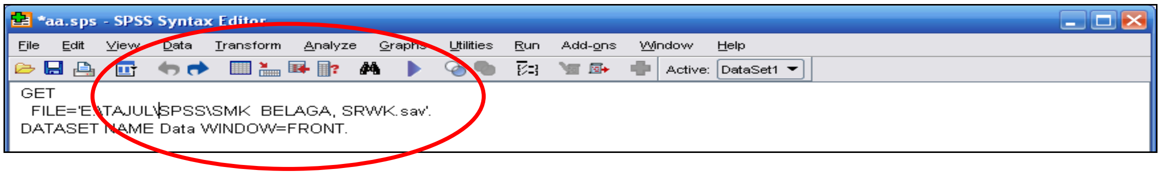
FIGURE 8: Data set after converted in SPSS 16.0

5. RESULT AND FINDING

This section will illustrate the analysis and result gained from the data collected. The descriptive analysis has been carried out to get some information from the outcomes of each prediction variable aligned with the targeted output.

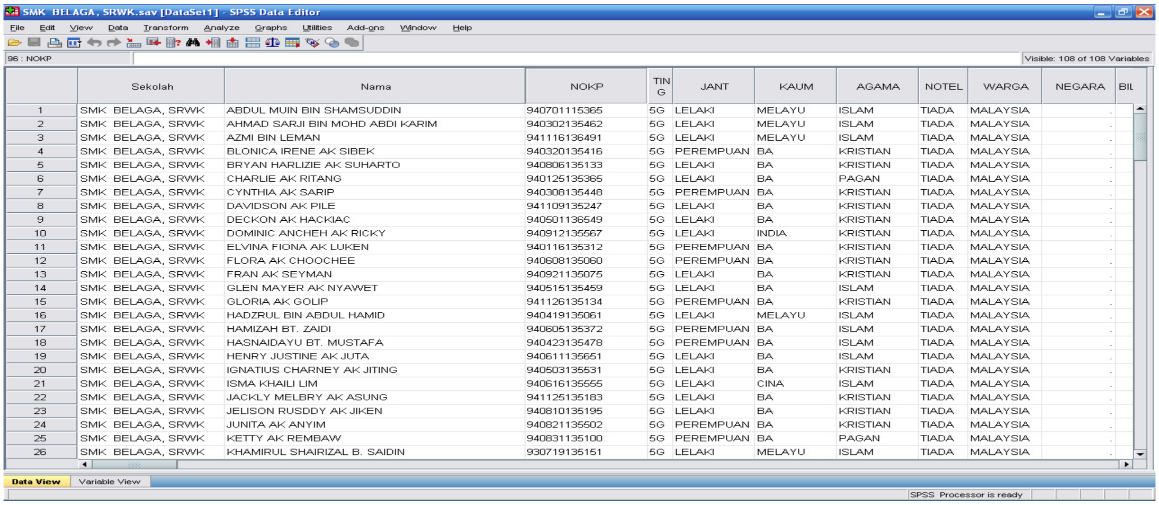
5.1 Data Integration

This study has illustrate how the process of data integration on data set has been applied for make it for the data analysis part. As you know, if the process of data integration has some error or fail, it will lead to the wrong analysis and wrong result. This phase is so important and it will be used SPSS 16.0 (Syntax) instrument to make this process done. There are seven data set which from technical schools in Malaysia has been selected as respondent data and need to integrated all together include SMK BELAGA SARAWAK, SMK INDERAPURA PAHANG, SMK KEPALA BATAS KEDAH, SMK KUALA KETIL KEDAH, SMK MARANG TERENGGANU, SMK SAMA GAGAH PULAU PINANG, and SMK TENGKU IDRIS SELANGOR. SPSS 16.0 software can be used manually to integrated these data but for this study purpose, it will be used Syntax editor in SPSS 16.0 to make it automated integration for these data. The original sources of dataset are in Microsoft Excel format and it will be imported to SPSS 16.0 program by using Syntax editor. The process of integration will be show below.



```
GET
  FILE='E:\TAJUL\SPSS\SMK  BELAGA, SRWK.sav'.
DATASET NAME Data WINDOW=FRONT.
```

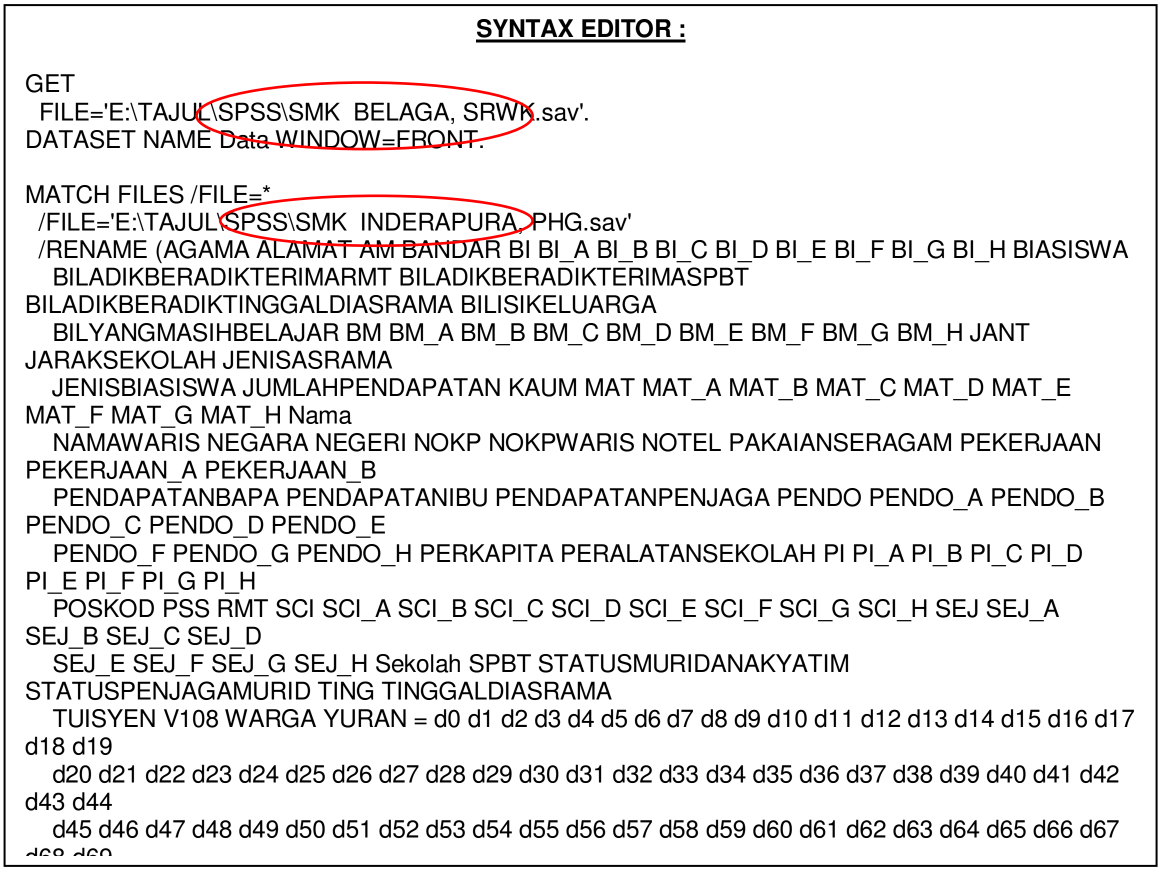
FIGURE 9: Source code in syntax editor to import data into SPSS 16.0



	Sekolah	Nama	NOKP	TING	JANT	KAUM	AGAMA	NOTEL	WARGA	NEGARA	BIL
1	SMK BELAGA, SRWK	ABDUL MUIN BIN SHAMSUDDIN	940701115365	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA	.	
2	SMK BELAGA, SRWK	AHMAD SARJI BIN MOHD ABDI KARIM	940302135462	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA	.	
3	SMK BELAGA, SRWK	AZMI BIN LEMAN	941116136491	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA	.	
4	SMK BELAGA, SRWK	BLONICA IRENE AK SIBEK	940320135416	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
5	SMK BELAGA, SRWK	BRYAN HARLIZIE AK SUHARTO	940806135133	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
6	SMK BELAGA, SRWK	CHARLIE AK RITANG	940125135365	5G	LELAKI	BA	PAGAN	TIADA	MALAYSIA	.	
7	SMK BELAGA, SRWK	CYNTHIA AK SARIP	940308135448	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
8	SMK BELAGA, SRWK	DAVIDSON AK PILE	941109135247	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
9	SMK BELAGA, SRWK	DECKON AK HACKIAC	940501136549	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
10	SMK BELAGA, SRWK	DOMINIC ANCHEH AK RICKY	940912135567	5G	LELAKI	INDIA	KRISTIAN	TIADA	MALAYSIA	.	
11	SMK BELAGA, SRWK	ELVINA FIONA AK LUKEN	940116135312	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
12	SMK BELAGA, SRWK	FLORA AK CHOOCHEE	940608135060	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
13	SMK BELAGA, SRWK	FRAN AK SEYMAN	940921135075	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
14	SMK BELAGA, SRWK	GLEN MAYER AK NYAWET	940515135459	5G	LELAKI	BA	ISLAM	TIADA	MALAYSIA	.	
15	SMK BELAGA, SRWK	GLORIA AK GOLIP	941126135134	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
16	SMK BELAGA, SRWK	HADZRUL BIN ABDUL HAMID	940419135061	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA	.	
17	SMK BELAGA, SRWK	HAMIZAH BT. ZAIDI	940605135372	5G	PEREMPUAN	BA	ISLAM	TIADA	MALAYSIA	.	
18	SMK BELAGA, SRWK	HASNAIDAYU BT. MUSTAFA	940423135478	5G	PEREMPUAN	BA	ISLAM	TIADA	MALAYSIA	.	
19	SMK BELAGA, SRWK	HENRY JUSTINE AK JUTA	940611135651	5G	LELAKI	BA	ISLAM	TIADA	MALAYSIA	.	
20	SMK BELAGA, SRWK	IGNATIUS CHARNEY AK JITING	940503135531	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
21	SMK BELAGA, SRWK	ISMA KHAILI LIM	940616135555	5G	LELAKI	CINA	ISLAM	TIADA	MALAYSIA	.	
22	SMK BELAGA, SRWK	JACKLY MELBRY AK ASUNG	941125135183	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
23	SMK BELAGA, SRWK	JELISON RUSDDY AK JIKEN	940810135195	5G	LELAKI	BA	KRISTIAN	TIADA	MALAYSIA	.	
24	SMK BELAGA, SRWK	JUNITA AK ANYIM	940821135502	5G	PEREMPUAN	BA	KRISTIAN	TIADA	MALAYSIA	.	
25	SMK BELAGA, SRWK	KETTY AK REMBAW	940831135100	5G	PEREMPUAN	BA	PAGAN	TIADA	MALAYSIA	.	
26	SMK BELAGA, SRWK	KHAMIRUL SHAIRIZAL B. SAIDIN	930719135151	5G	LELAKI	MELAYU	ISLAM	TIADA	MALAYSIA	.	

FIGURE 10: Dataset in SPSS 16.0 after import the data using syntax editor

Fig. 9 and 10 above have show how the dataset from Microsoft Excel format will be imported using SPSS syntax and will produce the output in SPSS 16.0 format by automatically without need used the manual from SPSS .

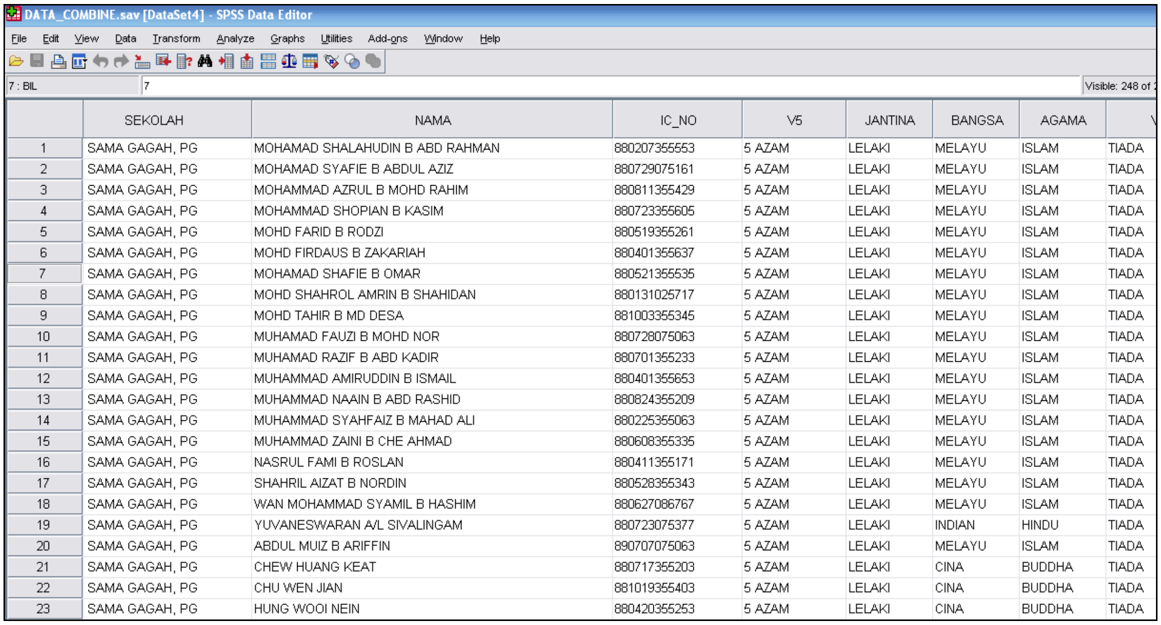


SYNTAX EDITOR :

```
GET
  FILE='E:\TAJUL\SPSS\SMK  BELAGA, SRWK.sav'.
DATASET NAME Data WINDOW=FRONT.

MATCH FILES /FILE=*
  /FILE='E:\TAJUL\SPSS\SMK  INDERAPURA, PHG.sav'
  /RENAME (AGAMA ALAMAT AM BANDAR BI BI_A BI_B BI_C BI_D BI_E BI_F BI_G BI_H BIASISWA
    BILADIKBERADIKTERIMARMT BILADIKBERADIKTERIMASPBT
BILADIKBERADIKTINGGALDIASRAMA BILISIKELUARGA
    BILYANGMASIHBELAJAR BM BM_A BM_B BM_C BM_D BM_E BM_F BM_G BM_H JANT
JARAKSEKOLAH JENISASRAMA
    JENISBIASISWA JUMLAHPENDAPATAN KAUM MAT MAT_A MAT_B MAT_C MAT_D MAT_E
MAT_F MAT_G MAT_H Nama
    NAMAWARIS NEGARA NEGERI NOKP NOKPWARIS NOTEL PAKAIANSERAGAM PEKERJAAN
PEKERJAAN_A PEKERJAAN_B
    PENDAPATANBAPA PENDAPATANIBU PENDAPATANPENJAGA PENDO PENDO_A PENDO_B
PENDO_C PENDO_D PENDO_E
    PENDO_F PENDO_G PENDO_H PERKAPITA PERALATANSEKOLAH PI PI_A PI_B PI_C PI_D
PI_E PI_F PI_G PI_H
    POSKOD PSS RMT SCI SCI_A SCI_B SCI_C SCI_D SCI_E SCI_F SCI_G SCI_H SEJ SEJ_A
SEJ_B SEJ_C SEJ_D
    SEJ_E SEJ_F SEJ_G SEJ_H Sekolah SPBT STATUSMURIDANAKYATIM
STATUSPENJAGAMURID TING TINGGALDIASRAMA
    TUISYEN V108 WARGA YURAN = d0 d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16 d17
d18 d19
    d20 d21 d22 d23 d24 d25 d26 d27 d28 d29 d30 d31 d32 d33 d34 d35 d36 d37 d38 d39 d40 d41 d42
d43 d44
    d45 d46 d47 d48 d49 d50 d51 d52 d53 d54 d55 d56 d57 d58 d59 d60 d61 d62 d63 d64 d65 d66 d67
d68 d69
```

FIGURE 11: Source code in syntax editor to make integration of all the respondent data

Fig. 11 above is some the source code in syntax SPSS to make integration of all these dataset by automatically control in syntax SPSS. Fig 11 above not include all the source code in syntax SPSS because of the space here. So it just can show four technical schools that will integrate together and the output will be save to another file in SPSS format which is 'DATA_COMBINE' as show in Fig. 12 below.



	SEKOLAH	NAMA	IC_NO	V5	JANTINA	BANGSA	AGAMA	V
1	SAMA GAGAH, PG	MOHAMAD SHALAHUDIN B ABD RAHMAN	880207355553	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
2	SAMA GAGAH, PG	MOHAMAD SYAFIE B ABDUL AZIZ	880729075161	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
3	SAMA GAGAH, PG	MOHAMMAD AZRUL B MOHD RAHIM	880811355429	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
4	SAMA GAGAH, PG	MOHAMMAD SHOPIAN B KASIM	880723355605	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
5	SAMA GAGAH, PG	MOHD FARID B RODZI	880519355261	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
6	SAMA GAGAH, PG	MOHD FIRDAUS B ZAKARIAH	880401355637	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
7	SAMA GAGAH, PG	MOHAMAD SHAFIE B OMAR	880521355535	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
8	SAMA GAGAH, PG	MOHD SHAHROL AMRIN B SHAHIDAN	880131025717	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
9	SAMA GAGAH, PG	MOHD TAHIR B MD DESA	881003355345	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
10	SAMA GAGAH, PG	MUHAMAD FAUZI B MOHD NOR	880728075063	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
11	SAMA GAGAH, PG	MUHAMAD RAZIF B ABD KADIR	880701355233	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
12	SAMA GAGAH, PG	MUHAMMAD AMIRUDDIN B ISMAIL	880401355653	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
13	SAMA GAGAH, PG	MUHAMMAD NAAIN B ABD RASHID	880824355209	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
14	SAMA GAGAH, PG	MUHAMMAD SYAHFAIZ B MAHAD ALI	880225355063	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
15	SAMA GAGAH, PG	MUHAMMAD ZAINI B CHE AHMAD	880608355335	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
16	SAMA GAGAH, PG	NASRUL FAMI B ROSLAN	880411355171	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
17	SAMA GAGAH, PG	SHAHRIL AIZAT B NORDIN	880528355343	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
18	SAMA GAGAH, PG	WAN MOHAMMAD SYAMIL B HASHIM	880627086767	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
19	SAMA GAGAH, PG	YUVANESWARAN A/L SIVALINGAM	880723075377	5 AZAM	LELAKI	INDIAN	HINDU	TIADA
20	SAMA GAGAH, PG	ABDUL MUIZ B ARIFFIN	890707075063	5 AZAM	LELAKI	MELAYU	ISLAM	TIADA
21	SAMA GAGAH, PG	CHEW HUANG KEAT	880717355203	5 AZAM	LELAKI	CINA	BUDDHA	TIADA
22	SAMA GAGAH, PG	CHU WEN JIAN	881019355403	5 AZAM	LELAKI	CINA	BUDDHA	TIADA
23	SAMA GAGAH, PG	HUNG WOOI NEIN	880420355253	5 AZAM	LELAKI	CINA	BUDDHA	TIADA

FIGURE 12: Output of data integration using Syntax SPSS.

5.2 Data Analysis

Data analysis for this study is involved of descriptive and predictive analysis that will be handling using Syntax SPSS code in SPSS 16.0. All the process in analysis part will be done by automatically using Syntax SPSS.

5.3 Descriptive Analysis

The collected data in this study has been processed using Syntax SPSS in SPSS version 16.0 to produce the experimental results. The analysis is divided into five main factors, such as number of family members, number of family member still learning, number of family member receive SPBT, family income and SPBT .

5.4 Number Of Family Members

The number of family members is one of the attribute of dataset that will be analyzed to see how it can influence the SPM result of student in technical schools. Initial analysis is done to make sure there is no missing value in the collected data. The analysis is shown below:

Statistics :	**Syntax SPSS Code :**
NUM_FAMILY_MEMBERS N Valid 691 Missing 0	GET FILE='E:\TAJUL\DATA\DATA_COMBINE.sav' . DATASET NAME DataSet1 WINDOW=FRONT. FREQUENCIES VARIABLES=NUM_FAMILY_MEMBERS /ORDER=ANALYSIS.

TABLE 6: Frequency data of Number of family members and source code in Syntax SPSS

Table 6 denotes there is no missing value for this attribute for used in the experiment and it was generate using Syntax SPSS code. In this attribute, there are 691 students whose number of family member's distribution is shown in Table 7.

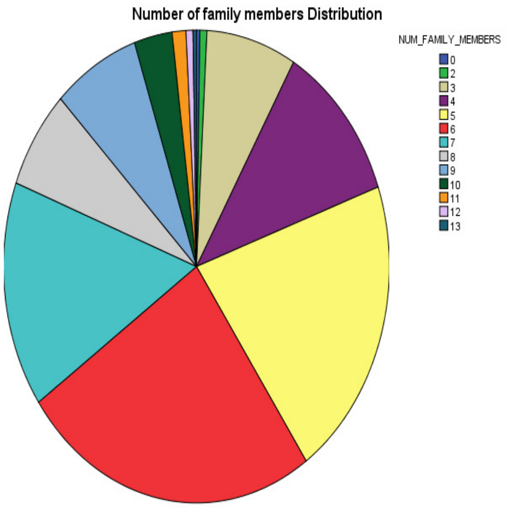
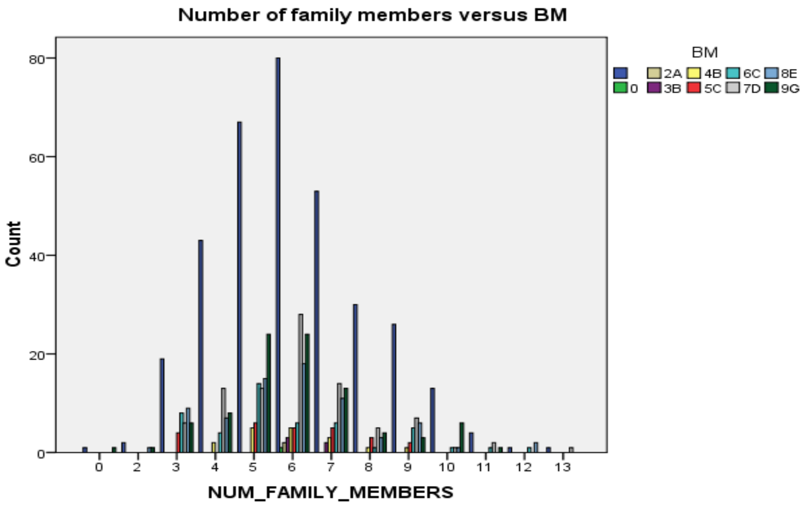
	**Syntax SPSS Code:**
	GGRAPH
	  /GRAPHDATASET NAME="graphdataset" VARIABLES=NUM_FAMILY_MEMBERS COUNT()[name="COUNT"]
	    MISSING=LISTWISE REPORTMISSING=NO
	  /GRAPHSPEC SOURCE=INLINE.
	BEGIN GPL
	  SOURCE: s=userSource(id("graphdataset"))
	  DATA: NUM_FAMILY_MEMBERS=col(source(s), name("NUM_FAMILY_MEMBERS"), unit.category())
	  DATA: COUNT=col(source(s), name("COUNT"))
	  COORD: polar.theta(startAngle(0))
	  GUIDE: axis(dim(1))
	  GUIDE: legend(aesthetic(aesthetic.color.interior), label("NUM_FAMILY_MEMBERS"))
	  SCALE: linear(dim(1), dataMinimum(), dataMaximum())
	  SCALE: cat(aesthetic(aesthetic.color.interior), sort.natural())
	  ELEMENT: interval.stack(position(summary.percent(summary.percent(COUNT,
	    base.all(acrossPanels()))))), color.interior(NUM_FAMILY_MEMBERS))
	END GPL.

TABLE 7: Number of family members

Table 7 has shown the number of family members from 0 to 13 persons. The highest percentage of the number of family members is 6 persons with 24.9% and the lowest percentages of the number of family members are 13 persons with 0.3%. The breakdown of the relationship between number of family members and each SPM's subject is evidently shown in Table 8.



Syntax SPSS :

```
CROSSTABS
  /TABLES=BM_SPM BY NUM_FAMILY_MEMBERS
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL
  /BARCHART.
```

TABLE 8: Number of family members versus BM

The number of family members of 6 persons shows their contribution in score with all grades for subject Bahasa Malaysia (BM). There are 2 students that get grade 2A, 3 students get grade 3B, 5 students get grade 4B, 5 students get grade 5C, 6 students get grade 6C, 28 students get grade 7D, 18 students get grade 8E and 24 students get grade 9G. The overall of this group of number family members of 6 persons was contribute for score in all grades are around 92 students. This process is automated generate using Syntax SPSS in SPSS 16.0.

This process need to continue for attributes number of family member still learning, number of family member receive SPBT, family income and SPBT. But it will not be show in this study because this study just wants to proof of focus on automated analysis that are automatically generate using Syntax SPSS and will be publish on web base data analysis result as show below (Fig 13 ~ Fig 14).

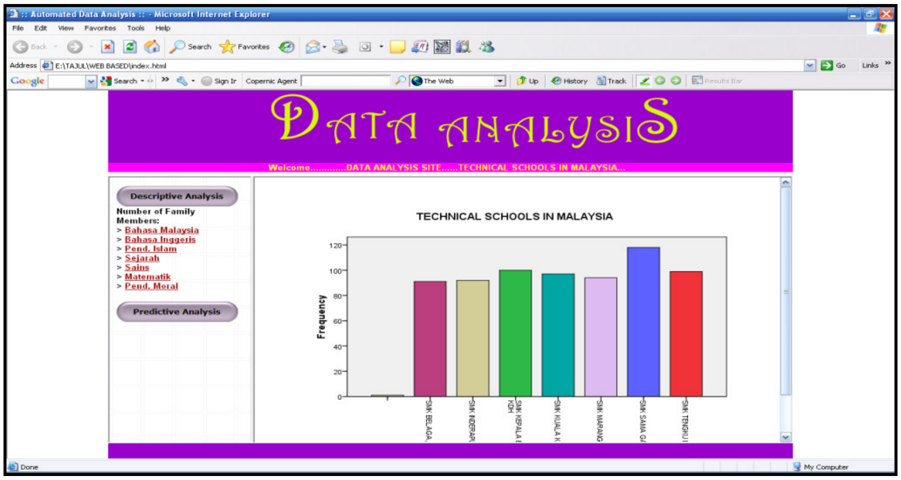


FIGURE 13: Main page on Data Analysis web based.

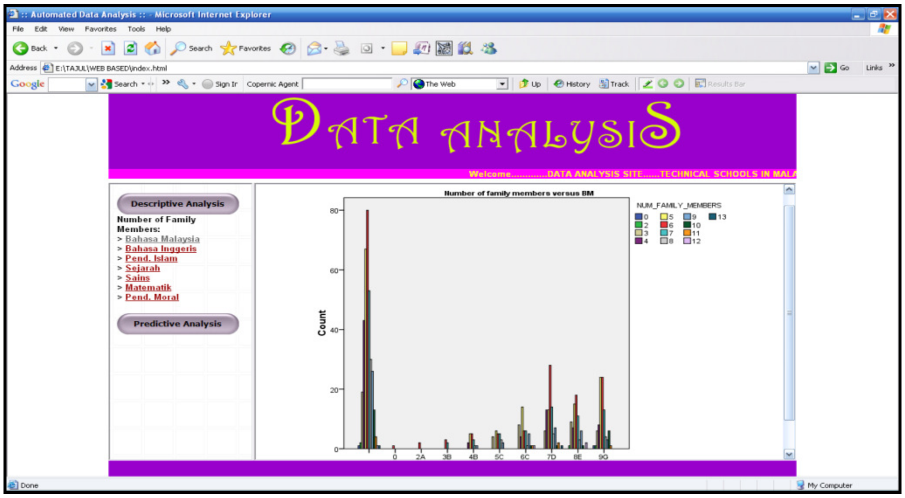

:: Automated Data Analysis :: - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address E:\TAJUL\WEB BASED\index.html
DATA ANALYSIS
Welcome............DATA ANALYSIS SITE......TECHNICAL SCHOOLS IN MALA
Descriptive Analysis
Number of Family Members:
> Bahasa Malaysia
> Bahasa Inggeris
> Pend. Islam
> Sejarah
> Sains
> Matematik
> Pend. Moral
Predictive Analysis
Number of family members versus BM
Count
0
20
40
60
80
0
2A
3B
4B
5C
6C
7D
8E
9G
NUM_FAMILY_MEMBERS
0
2
3
4
5
6
7
8
9
10
11
12
13
Done
My Computer

FIGURE 14: Description analysis on Number of family members versus BM

6. CONCLUSION

This study summarized the examination factors such as exam results and other factors such as SPM's subject, number of family members, number of family member still learning, number of family member receive SPBT, family income and SPBT that contribute to students' academic achievement in the future. These all factors show strong relationship between each other. The descriptive analysis was used to describe the frequency and cross tabulation between variables in this study.

As this study has illustrated, there is a potential of developing a system for centralizing the secondary technical schools student data. It also possible to perform automated descriptive analysis online that could reduce the time required to process the marks and perform manual analysis. The study also demonstrates that it is possible to integrate the proposed system with statistical analysis package in order to deliver intelligent business solutions.

Another finding from the study indicates that the analysis obtained could be used by the school management to make a suitable plan for their students' academic achievement program in the future. In addition, other data mining techniques like association rule also can be used to measure the association between attributes. The finding could be used to further enhance the strength of each attribute with description variables and among attributes or independent variables.

7. ACKNOWLEDGMENT

This presented study is supported by the Jabatan Pelajaran Negeri Pulau Pinang for their support direct and indirect in order to complete this research.

8. REFERENCES

[1] Terry E, Spradlin, Kirk R, Walcott C, Kloosterman P, Zaman K, McNabb S, Zapf J & associates, "Is The Achievement Gap in Indiana Narrowing", *Education Resources Information Center Journal,* September 2005.

[2] Cripps A, "Using Artificial Neural Nets to Predict Academic Performance," *American Psychological Association Journal*, pp. 33 – 37, Feb.1996.

[3] Beal, C. R. & Cohen, P. R. (2006). Temporal Data Mining for Educational Applications. Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

[4] Hayek, John C, Kuh, George D, "College Activities and Environmental Factors Associated with The Development of Life Long Learning Competencies of College Seniors" *Education Resources Information Center Journal,* November 1999.

[5] Henchey, Norman, "Schools That Make A Difference : Final Report. Twelve Canadian Secondary Schools in Low Income Settings" *Education Resources Information Center Journal,* November 2001.

[6] Gibson, Margaret A, "Improving Graduation Outcomes for Migrant Students", *Education Resources Information Center Journal,* July 2003.

[7] Ma, Y., Liu, B., Wong, C. K., Yu, P. S. & Lee, S. M. (2000). Targeting the Right Students Using Data Mining.

[8] Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). *Data mining for business intelligence : concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. Hoboken, NJ: John Wiley & Sons.

New Framework to Detect and Prevent Denial of Service Attack in Cloud Computing Environment

Mohd Nazri Ismail *mnazrii@miit.unikl.edu.my*
Malaysian Institute of Information Technology (MIIT)
University Kuala Lumpur,Malaysia

Abdulaziz Aborujilah *abdulazizh@unikl.edu.my*
Malaysian Institute of Information Technology (MIIT)
University Kuala Lumpur,Malaysia

Shahrulniza Musa *shahrulniza@miit.unikl.edu.my*
Malaysian Institute of Information Technology (MIIT)
University Kuala Lumpur,Malaysia

AAmir Shahzad *mail2aamirshahzad@gmail.com*
Malaysian Institute of Information Technology (MIIT)
University Kuala Lumpur,Malaysia

Abstract

As a result of integration of many techniques such as grading, clustering, utilization computing and resource's sharing, cloud computing has been appeared as multi element's composition technology, it offers several computing services such as IaaS (infrastructure as service), PaaS (platform as service) and SaaS (software as service) based on pay as you use rule, but nevertheless, and because of cloud computing end users participate in computing resources (co_ tenancy) , and by which infrastructure computing can be shared by a number of users, and as a result to this feature, some security challenges has been existed and one of the most serious security threats is flooding attack, which prevent other users from using cloud infrastructure services, that kind of attack can be done by a legitimate or illegitimate cloud computing users.

To overcome this problem various approaches have been proposed based on Artificial intelligence and statistical methods, but most of them concentrate on one side of problem and neglect the other aspects.

In our proposed approach, the focusing will be more in overcoming the problem in all its aspects, in attack detection stage covariance matrix statistical method will be applied and to determine attack source TTl (Time_to_Life) value counting method will be used, and the attack prevention will be based on Honeypot method, and initial simulation to this approach using UML class diagram and sequence diagram showed where our proposed framework can be done in cloud environment.

Keywords: Flooding based denial-of-service (DDoS) attack, Covariance matrix, TTl, Honeypot, cloud computing, virtual machine.

1. INTRODUCTION

Because of current powerful computing capabilities (CPU, memory, storage media) and also as a result to networking capabilities such as grade [1] and cluster computing [2] and due to vitalization techniques [3] as in Figure 1, cloud computing services have taken a place in modern computing technology. Cloud computing can be defined as in [4] “Cloud computing is TCP/IP based high development and integrations of computer technologies such as fast microprocessor, huge memory, high-speed network and reliable system architecture."

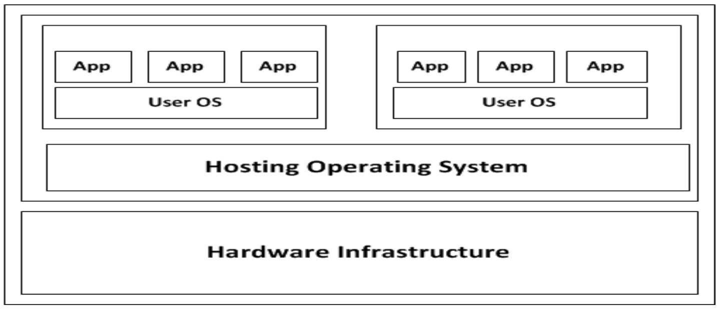


FIGURE 1: cloud computing virtualization

In this paper, we first present a general background about cloud computing architecture after that cloud computing challenges will be highlighted, then we will start in literature review and related works which concentrate on DoS detecting and prevention method, finally we describe our proposed defense framework and initial simulation, we use flooding based attack and DoS attack interchangeably .

2. CLOUD COMPUTING ARCHITECTURE

The cloud computing as new technology help the end user to reducing required computing effort to achieve his goals .and because of cloud computing services provided to the user such as IaaS (infrastructure as a service) [5], the end user is no longer needs to purchase computer equipment necessary to accomplish his goal, but he can rent all equipment he needs then use them as much as he needs.

In addition to IaaS provided by cloud computing providers, computer application developers can take advantage of the multiple development platforms available on the cloud to develop their own applications and deployed them online, which is known as PaaS(platform as the service) [6] .
In addition to IaaS and PaaS services, the end user can use several application's software in different fields available on cloud by so-called SaaS (software as services)[7], and cloud computing application software provided by main cloud providers can reduce the cost of rent or purchase those applications to the end user.

Although these services offered by cloud computing provides, but the most important shortage in cloud services is the security side which is mentioned in details in [8-11], and one of the most serious security problem in cloud computing is related to availability of cloud infrastructure, because of many cloud user can share the same infrastructure, that can cause security threat such as denial of service attack [10-16].

3.CLOUD COMPUTING CHALLENGES

Cloud computing implementation is facing several challenges in different aspect [17] and According to the survey conducted in 2008 by IDC survey about cloud computing challenges, cloud computing security challenges is on the top most threat to cloud [18], and became well known that cloud computing paradigm is a kind of virtualization environment with another technology such as grade and clusters and distributed computing, all these technologies has his own security disadvantages. In addition to the threats coming from cloud component, therefore securing issues is not related to cloud just but also related to other technologies and most dangers threat to cloud is vitalization security.

And one of the potential attacks to cloud virtualization system is neighbor attacks as in Figure 2, which by any virtual machine can attack its neighbor in same physical infrastructures and thus prevent it from providing its services or, which has been known as denial of service attack DoS attack as has been existed in AWS Amazon [19], that kind of attack can effect on cloud performance in general and can cause financial Losses [20] and can cause harmful effect in other servers in same cloud infrastructure as in [10].

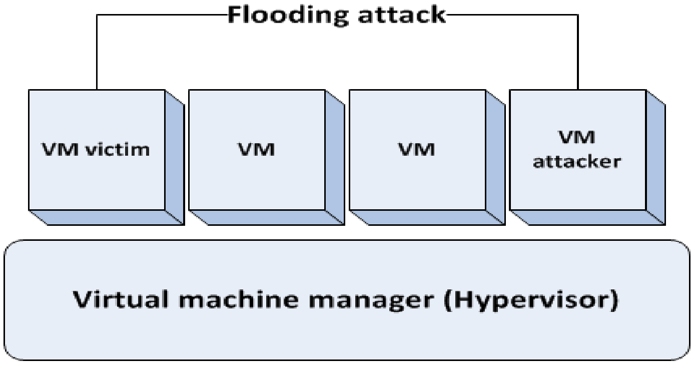


FIGURE 2: Flooding Dos attack in cloud environment

4. LITERATURE REVIEW and RELATED WORK

Denial of service attack poses as one of the most networks famous attacks, by which one victim machine can receive more than its capacity and so other end users requests cannot be served by server, and in the cloud environment, that kind of attack can be most harmful than unclouded environment because of VMs neighboring and resource sharing in cloud computing environment, so one virtual machine can be used as a source of denial of service attack to another virtual machine in same infrastructure and for overcome this security threat , several kind of flooding DoS attacks detecting and prevention approach has been suggested, and every one propose a method to detect or prevent this attack .

4.1 DETECTNG METHODS

Several kinds of flooding DoS attacks detecting approach has been suggested [21], One of these methods has proposed a kind of pattern generation for spatial-temporal traffic pattern in the application layer according to document popularity and also Access Matrix and behavior of web access. This method depends on semi-Markov modeling, and potential DoS attack is identified by entropy of document popularity, which matching the model [22].

In [23] The study pays more attention to service violations as an indicator to DoS attack, then the author explains comparison study to different kind of network monitoring techniques in terms of overhead, finally some of the parameters have been proposed to help in choosing the best monitoring schema according to the requirements and amount of overhead and Permittivity.

In [24] the author proposed framework to detect and trace back the source of attack, the first stage could detect two kinds of attack, logical and DoS/DDoS ,Neural networks used to discover logical attacks while DoS/DDoS can be recognized by CUSUM algorithm and for track backing attack, sources hash tracer is used then the second component of framework is prediction model which focuses on recognize of malicious behaviors for network traffic .

In [25] this research queuing theory model has been suggested to detect DoS attack on line, the proposed approach depends on two models , the first is to detect abrupt change based on some of the parameter and second model is the signal generation module which used for further processing.

In [26] the author has studied the dependency of web page attributes to detect new kinds of application layer attack then multiple principal component analysis is adapted as modeling method to model normal web browsing behaviors and finally, the author conducted some experiments to prove his model.

In [27] the author has developed a simulation to forecast numbers of zombies used in DDoS attack ,the forecasting depend on a polynomial regression model, and several statistical measures have been used for performance evaluation, and NS2 has been utilized as simulation platform and the result has been proven that this detection approach can forecast the number of zombies used in DDoS attack efficiently with low error percentage and prediction number of zombies in a DDoS Attack using Polynomial Regression Model.

in[28] this research paper, packet flows analysis for some network protocol has been suggested to detect DDoS attack and for normal behavior forecasting Gaussian parametric mixture has been utilized finally and for detect DDoS attack queue approach has been implemented, and the results showed that these detecting methods have accepted.

In [29] the author suggest using several computers working together hosted in a cloud to monitoring and analysis network traffic in same time and identify potential attack and for achieving that internet network can be used to link all detection devices systems together.

in [21] covariance-matrix statistical approach has been used for flooding based DoS attack detection, covariance-matrix depend on study and monitor of network traffic features correlativity changes and compare the covariance matrix of normal traffic and any new observed traffic and classify the comparison results according predefined threshold and finding the degree of anomaly of new captured traffic and normal traffic profile, and implementation of this approach has proven more accuracy and efficiency through simulation experiments to two of most famous flooding based attack Neptune and Smurf attacks , and in addition to high accuracy to this approach ,it can also detect second order of features which can be possible attack finally covariance-matrix can be summarized in three spaces, captured traffic space ,which content TCP dump data format and covariance-matrix space and lastly decision space which decide about the traffic either attack traffic or normal traffic as in Figure 3

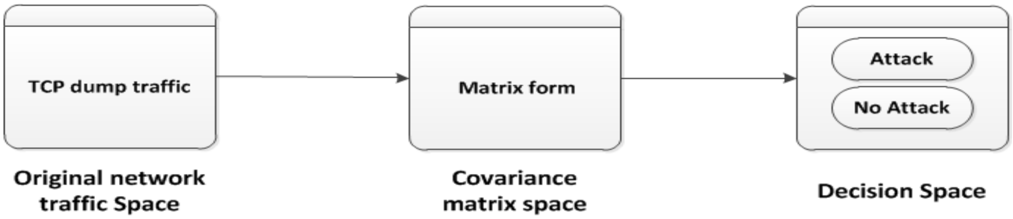


FIGURE 3: Covariance matrix detecting approach spaces

In [30] The author suggested framework depend on sharing IDS warring alerts with each other by set up IDS client tool and estimate the most dangerous attack according to the number of voting obtained by IDS client and simulation of this model has been done in snore environment, and the result has shown that this approach can protect a network from one point of failed

4.2 Prevention Methods

In [31] the author suggests a new approach to mitigate of flooding attack based on users' behavior's history using hidden markov modeling after traffic filtering process and also the author determined the implementation requirement such as firewall and access control setting, and for prove effectiveness of this approach DDoS simulation has been done .

In [32] the author proposes connected firewall to internet trap attack packets before its arrival to its victim. In [14]the author of this paper suggest a prevention strategy to avoid DDoS attack that targets application by shifting any cloud application under attack to another virtual machine in another physical server.

In [33] the author analysis DDoS attacks in random flow network environment, he suggests using this model to evaluate DDoS prevention frameworks, and the simulation has shown the relationship between multi matrixes inferred from the model.

In [34] research paper author suggested a new congestion control mechanism in computer networks to protect it from DDoS attack, this approach depended mainly on filtering mechanism based on time factor, and the author proposed using this protection method before applying queue management rules, and performance of this method using NS2 has done according to IP traces reported in http://www.nlnar.org

In [35] this paper proposed a bandwidth based DDoS prevention method by classify the traffic into three kinds, first is normal, attack and suspicious traffic, attack traffic is restrictive and the if the victim confirmed that arrived traffic is suspicious, then it determines the source of attack to block attack traffic coming from.

In [36] the requirements of the DDoS attack prevention system have been suggested, and the author suggested also DDoS attack prevention platform integration process

In [36] the author suggest finding and prevent TCP Spoofed IP address attack to mitigate TCP SYN flooding attack through TCP specific probing procedure, which forces the end user to modify windows size in packet retransmission through three handshake process in TCP protocol , previous work in mitigate TCP SYN flooding rely on count of SYN flag sent by same IP, but it was weak in order to preventing this attack because attackers using Spoofed IP so needing to this way of preventing TCP Spoofed IP address attack is very necessary

In [37] the author of this paper suggests three procedures to reduce TCP syn attack effects, firstly, utilized router as DoS defense then reduces SYN attack effects by blocking TCP, involving Trusted Platform Module in network infrastructure and Finlay using Certified System Defense.

5. PROPOSED FRAMEWORK

Proposed framework depends on covariance matrix mathematical modeling with three stages, firstly training stage then detecting and prevention stages as the following:

5.1 Training Stage

The first stage is monitoring income network traffic in virtual switch using any flow traffic tool such as wireshark[38] or snort [39] , in normal case or without attack, the traffic behavior has normal data distribution but with attack traffic data distribution change into abnormal form, this assumption can be as rule in detecting process.

First stage in detecting stage is summarizing all packets traffic in matrix values form, after that matrix is converted into a covariance matrix (normal traffic profile), finally resulted matrix is stored for further analysis as in Figure 4.

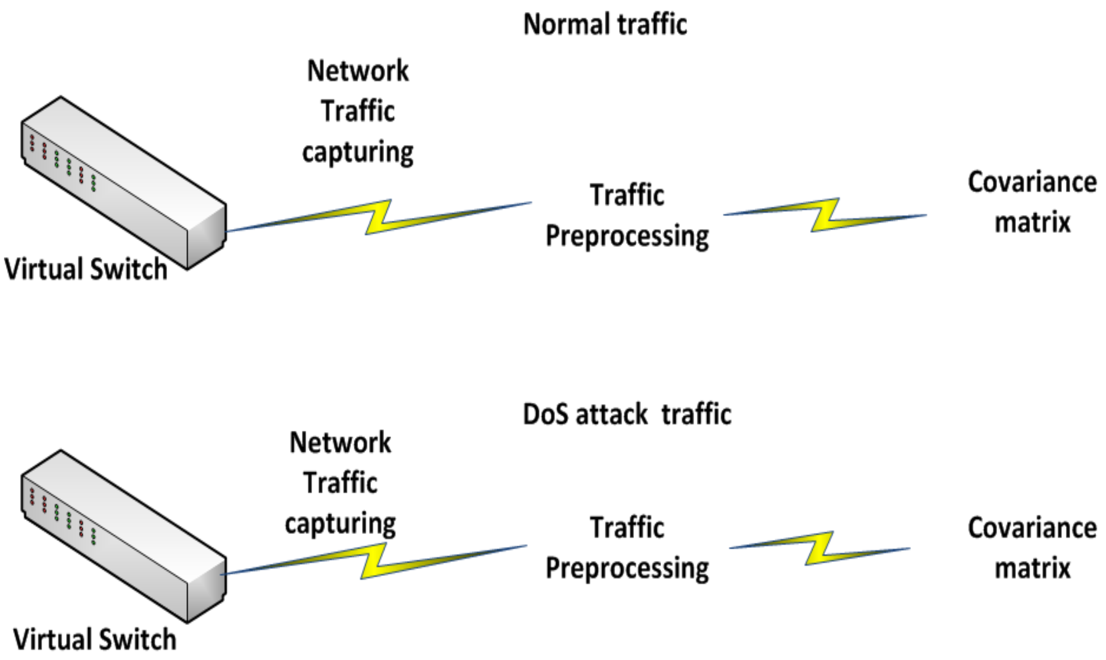


FIGURE 4: DoS attack detecting (Training stage)

5.2 DETECTING AND PREVENATOIN STAGES

In these stage, covariance matrix resulted from new captured traffic is compared with profile of normal traffic(covariance matrix of normal traffic), whenever the result matrix is all zero's values that mean no attack and whenever the resulted matrix not all zeros value and the anomaly degree values more than a predefined threshold, that mean attack has happened and thus detecting signal appears and focus will move into the stage of network protection by finding out attack source IP address.

And to find attack source IP address, number of nodes that attacker pass through until reach victim side is counted by counting value of TTL (Time _to_Life)[40]. After determining the source of attack, all IP address used by an attacker is blocked using honypot network which pings all IP address used by attacker and whenever there is a replay, all responded IP address is blocked [14].

Then finally when the attack has been known, the legitimate traffic to victim's virtual machine is shifted to same virtual machine but in another physical machine [14], because cloud computing environment located multiple copies for one virtual machine to strengthening reliability of computing as in Figure 5

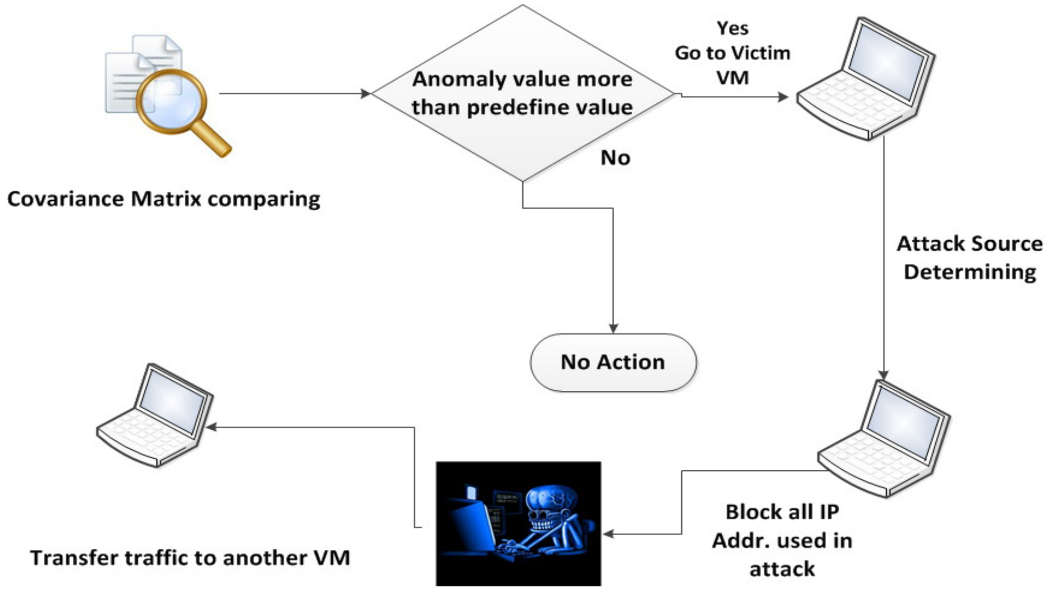

Covariance Matrix comparing
Anomaly value more than predefine value
Yes
Go to Victim VM
No
No Action
Attack Source Determining
Block all IP Addr. used in attack
Transfer traffic to another VM

FIGURE 5: DoS attack prevention stage

6. PROPOSED FRAMEWORK TEST BED

To implement the proposed framework, simple cloud environment can be built, which consists of three web server virtual machine and user can access to web server by internet gateway and virtual switch as in Figure 6

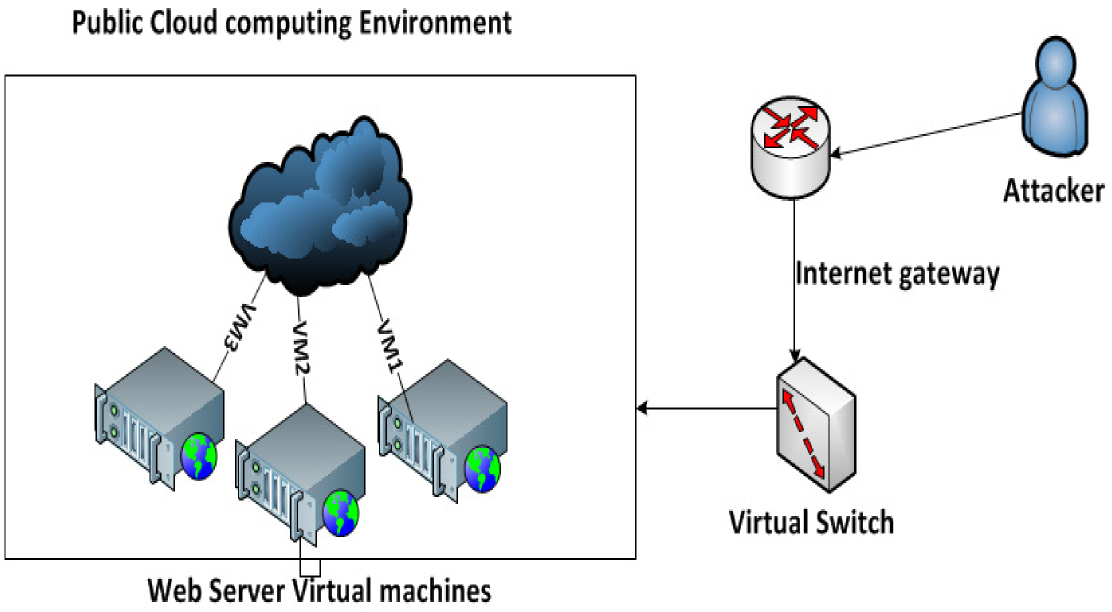

Public Cloud computing Environment
VM3
VM2
VM1
Web Server Virtual machines
Attacker
Internet gateway
Virtual Switch

FIGURE 6: Framework experiment test bed

7. IMPLEMENTATION

Proposed framework can be simulated Initially in conceptual level , using class and sequences UML diagrams [41] and Covariance mathematical model can be involved in cloud environment simulating to detect flooding DoS attack as in Figure 7 and Figure 8 and all detailed of simulation in [40].

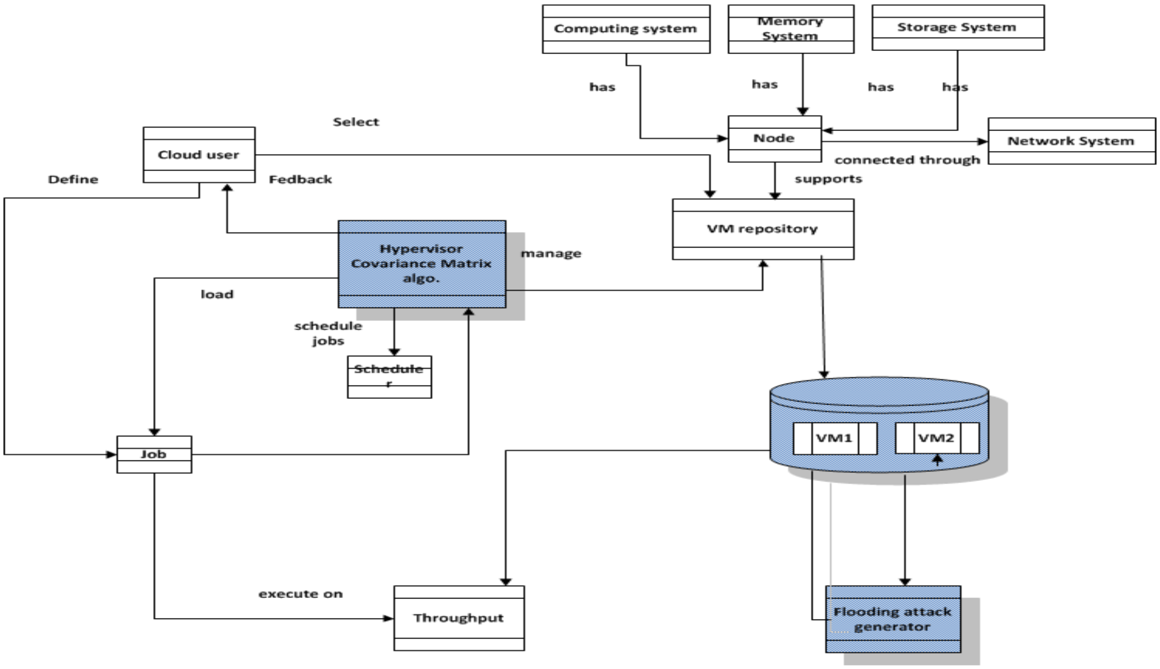


FIGURE 7: Class diagram to cloud environment

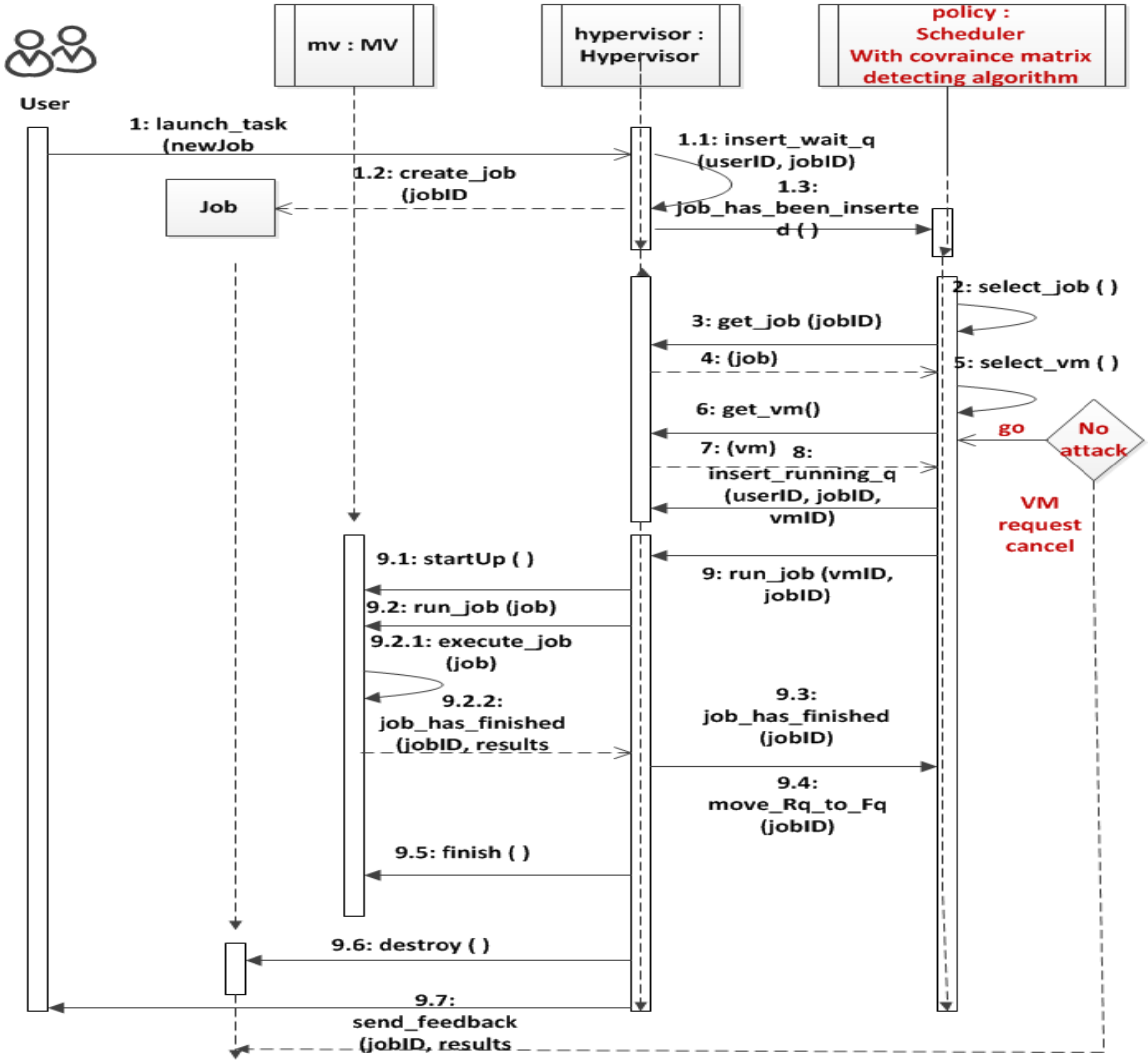


FIGURE 8: Sequence diagram to cloud environment

8. CONCLUSION

Finlay we can conclude that as cloud computing can offer new computing benefits, but it faces high risks, specifically in the security side where DDoS attack can make cloud service unavailable, and several methods have suggested but most of them give more attention to one side either detecting or track backing or prevention, our new frame work focuses on all aspects of the problem. Simulation part we give a conceptual view to location of the covariance matrix in cloud computing environment using class and sequences UML diagram and In future work we plan to performance the simulation mentioned above and implement the proposed frame work in real cloud environment and determine some constraints such as on line detecting constraints and also detecting the attack in different cloud environment such as private, public and hyper environment.

9. REFERENCE

[1] Foster, I. and C. Kesselman, The grid: blueprint for a new computing infrastructure. 2004: Morgan Kaufmann.

[2] Buyya, R., High performance cluster computing: programming and applications, vol. 2. Pre ticeHallPTR, NJ, 1999.

[3] Armbrust, M., et al., A view of cloud computing. Communications of the ACM, 2010. 53(4): p. 50-58.

[4] Mell, P. and T. Grance, The NIST definition of cloud computing. National Institute of Standards and Technology, 2009. 53(6): p. 50.

[5] Bhardwaj, S., L. Jain, and S. Jain, Cloud computing: A study of infrastructure as a service (IAAS). International Journal of engineering and information Technology, 2010. 2(1): p. 60-63.

[6] Kulkarni, G., P. Khatawkar, and J. Gambhir, Cloud Computing-Platform as Service. International Journal of Engineering. 1.

[7] Kulkarni, G., J. Gambhir, and R. Palwe, Cloud Computing-Software as Service. International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012. 1(1).

[8] Foster, I., et al. Cloud computing and grid computing 360-degree compared. 2008: Ieee.

[9] Ngongang, G., Cloud Computing Security. 2011.

[10] Subashini, S. and V. Kavitha, A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 2011. 34(1): p. 1-11.

[11] Chen, Y., V. Paxson, and R.H. Katz, What's new about cloud computing security? University of California, Berkeley Report No. UCB/EECS-2010-5 January, 2010. 20(2010): p. 2010-5.

[12] Chonka, A., et al., Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks. Journal of Network and Computer Applications, 2010.

[13] Hwang, K., S. Kulkareni, and Y. Hu. Cloud security with virtualized defense and reputation-based trust mangement. 2009: Ieee.

[14] Bakshi, A. and B. Yogesh. Securing cloud from ddos attacks using intrusion detection system in virtual machine. 2010: Ieee.

[15] Almulla, S.A. and C.Y. Yeun. Cloud computing security management. 2010: IEEE.

[16] Iankoulova, I. and M. Daneva, Cloud Computing Security Requirements: a Systematic Review.

[17] Bamiah, M.A. and S.N. Brohi, Seven Deadly Threats and Vulnerabilities in Cloud Computing.

[18] Dillon, T., C. Wu, and E. Chang. Cloud computing: Issues and challenges. 2010: Ieee.

[19] Hovav, A. and J. D'Arcy, The Impact of Denial□of□Service Attack Announcements on the Market Value of Firms. Risk Management and Insurance Review, 2003. 6(2): p. 97-121.

[20] Peng, T., C. Leckie, and K. Ramamohanarao, Survey of network-based defense mechanisms countering the DoS and DDoS problems. ACM Computing Surveys (CSUR), 2007. 39(1): p. 3.

[21] Yeung, D.S., S. Jin, and X. Wang, Covariance-matrix modeling and detecting various flooding attacks. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2007. 37(2): p. 157-169.

[22] Xie, Y. and S.Z. Yu, Monitoring the application-layer DDoS attacks for popular websites. Networking, IEEE/ACM Transactions on, 2009. 17(1): p. 15-25.

[23] Habib, A., M. Hefeeda, and B. Bhargava. Detecting service violations and DoS attacks. 2003.

[24] Leu, F., Intrusion Detection, Forecast and Traceback Against DDoS Attacks. 2009.

[25] Singh, N., S. Ghrera, and P. Chaudhuri, Denial of Service Attack: Analysis of Network Traffic Anormaly using Queuing Theory. Arxiv preprint arXiv:1006.2807, 2010.

[26] Lee, S., G. Kim, and S. Kim, Sequence-order-independent network profiling for detecting application layer DDoS attacks. EURASIP Journal on Wireless Communications and Networking, 2011. 2011(1): p. 1-9.

[27] Gupta, B., R. Joshi, and M. Misra, Prediction of Number of Zombies in a DDoS Attack using Polynomial Regression Model. Journal of Advances in Information Technology, 2011. 2(1): p. 57-62.

[28] Hao, S., et al. A queue model to detect DDos attacks. 2005: IEEE.

[29] Guilbault, N. and R. Guha. Experiment setup for temporal distributed intrusion detection system on amazon's elastic compute cloud. 2009: IEEE.

[30] Lo, C.C., C.C. Huang, and J. Ku. A cooperative intrusion detection system framework for cloud computing Networks. 2010: IEEE.

[31] Prabha, S. and R. Anitha, Mitigation of Application Traffic DDOS Attacks with Trust and Am Based Hmm Models. International Journal of Computer Applications IJCA, 2010. 6(9): p. 26-34.

[32] Chang, R.K.C., Defending against flooding-based distributed denial-of-service attacks: A tutorial. Communications Magazine, IEEE, 2002. 40(10): p. 42-51.

[33] Kong, J., et al. Random flow network modeling and simulations for DDoS attack mitigation. 2003: IEEE.

[34] Hu, Y.H., H. Choi, and H.A. Choi. Packet filtering to defend flooding-based DDoS attacks [Internet denial-of-service attacks]. 2004: IEEE.

[35] Wuu, L.C., et al. A practice of the intrusion prevention system. 2007: IEEE.

[36] Choi, Y.S., et al. Integrated DDoS attack defense infrastructure for effective attack prevention. 2010: IEEE.

[37] Chao-yang, Z. DOS Attack Analysis and Study of New Measures to Prevent. 2011: IEEE.

[38] Lamping, U. and E. Warnicke, Wireshark User's Guide. Interface, 2004. 4: p. 6.

[39] Roesch, M. Snort-lightweight intrusion detection for networks. 1999: Seattle, Washington.

[40] Wang, H., C. Jin, and K.G. Shin, Defense against spoofed IP traffic using hop-count filtering. IEEE/ACM Transactions on Networking (TON), 2007. 15(1): p. 40-53.

[41] Nunez, A., et al. Design of a flexible and scalable hypervisor module for simulating cloud computing environments. 2011: IEEE.

Software Design Level Vulnerability Classification Model

Shabana Rehman *shabana.infosec@gmail.com*
Department of Information System
Salman bin Abdul Aziz University
Al-Kharj, KSA

Khurram Mustafa *kmustafa@jmi.ac.in*
Department of Computer Science
Jamia Millia Islamia
New Delhi, 110025, India

Abstract

Classification of software security vulnerability no doubt facilitates the understanding of security-related information and accelerates vulnerability analysis. The lack of proper classification not only hinders its understanding but also renders the strategy of developing mitigation mechanism for clustered vulnerabilities. Now software developers and researchers are agreed on the fact that requirement and design phase of the software are the phases where security incorporation yields maximum benefits. In this paper we have attempted to design a classifier that can identify and classify design level vulnerabilities. In this classifier, first vulnerability classes are identified on the basis of well established security properties like authentication and authorization. Vulnerability training data is collected from various authentic sources like Common Weakness Enumeration (CWE), Common Vulnerabilities and Exposures (CVE) etc. From these databases only those vulnerabilities were included whose mitigation is possible at the design phase. Then this vulnerability data is pre-processed using various processes like text stemming, stop word removal, cases transformation. After pre-processing, SVM (Support Vector Machine) is used to classify vulnerabilities. Bootstrap validation is used to test and validate the classification process performed by the classifier. After training the classifier, a case study is conducted on NVD (National Vulnerability Database) design level vulnerabilities. Vulnerability analysis is done on the basis of classification result.

Keywords: Security Vulnerabilities, Design Phase, Classification, Machine Leaning, Security Properties

1. INTRODUCTION

Developing secure software remains a significant challenge for today's software developers as they still face difficulty in understanding the reasons of vulnerabilities in the existing software. It is vital to be able to identify software security vulnerabilities in the early phases of SDLC (Software Development Lifecycle) and one of early detection approaches is to consult with the prior known vulnerabilities and corresponding fixes [1]. Identification of candidate security vulnerability pays a substantial benefit when they are deals in early phases like requirement and design phases of the software [2]. Classification of vulnerabilities is fruitful in understanding the vulnerabilities better and classification also helps in mitigating group of vulnerabilities. Identifying and mitigating security vulnerabilities is no doubt a difficult task therefore taxonomy is developed that can classify vulnerabilities into classes and this will help designer to mitigate cluster of vulnerabilities. There are number of approaches of taxonomy development in past, like [3,4,5,6] etc, but no one ever propose any taxonomy that classify design level vulnerabilities on the basis of security properties. We have already proposed a taxonomy in [7], as shown in Table 1.0 (a) in which, priori classification is proposed and vulnerabilities are classified manually. But in this classification there is chance of 'Hawthorne Effect', it also largely depends on the expertise of the classifier. Therefore here we are creating a classifier that can classify a vulnerability data automatically. Machine learning is now a popular tool in the automation task. Researchers have explored the use of machine learning techniques to automatically associate documents with categories by first using a training set to adapt the classifier to the feature set of the particular document set [8]. Machine learning is a relatively new approach that can be used in classifying vulnerabilities. Therefore here Classifier is proposed, that classifies vulnerabilities on the basis of previously identified vulnerability and can help designer to place vulnerability in the predefined vulnerability classes that are based on the security properties of the software. Therefore mitigation mechanism can be applied for the whole class of vulnerabilities. In this classifier, first data pre-processing is done like text stemming, stop word removal, case transformation then SVM (Support Vector Machine) is used for the final classification with the regression model. Several conclusions are drawn after applying a classification. At last using this classifier NVD (National Vulnerability Database) vulnerabilities are classified and analyzed.

First Level	Second Level	Third level	Fourth Level
Access Control	Access Control at Process Level	Authentication	Missing Authentication procedure
			Insufficient Authentication procedure
			Wrong Authentication procedure
		Authorization	Missing Authorization procedure
			Insufficient Authorization procedure
			Wrong Authorization procedure
		Audit & logging	Missing Audit and logging
			Insufficient Logging or Audit of information
			Wrong Audit or Logging of information
	Access Control at Communication Level	Secured Session Management	Missing Secured Session management
			Insufficient Secured Session Management
			Wrong Secured Session Management
		Secured Information Flow	Missing Encryption of Sensitive Data During Transmission
			Insufficient Encryption of Sensitive Data during Transmission
			Wrong Encryption of Sensitive Data during Transmission
	Exposures leading to Access Violation	Exposures in Error Message	Missing Secured Error Message
			Insufficient Secured Error Message
			Wrong Secured Error Message
		Predictable Algorithm /sequence numbers/file names	Missing Randomness in the Random Sequence Ids
			Insufficient Randomness in the Random Sequence Ids
			Wrong Randomness in the Random Sequence Ids or Wrong Choice of File Name
		User Alertness	Missing User Alerting Information
			Insufficient User Alerting Information
			Wrong User Alerting Information

TABLE 1.0 (a): Taxonomy of Design Level Vulnerabilities

While considering the number of classes in the proposed classifier, we consider only ‘access control at process level’ and ‘access control at communication level’ and all the other type of vulnerabilities are considered in the ‘Others’ class. Because ‘exposure leading to access violation’ class covers a large domain of vulnerabilities and needs a separate study, therefore after exploring the domain of this class, we exclude this from the classifier and will consider for the future work.

Rest of the paper is organized as follows, in section 2, related works in the vulnerability classification is discussed, and then in section 3, the development process of vulnerabilities classification model is explained in detail. Classification of vulnerabilities using developed classifier is done in section 4. Conclusion and future work are discussed in section 5.

2. RELATED WORK

There are many classification approaches using machine learning techniques like [9] proposed uses a ontological approach to retrieving vulnerability data and establishing a relationship between them, they also reason about the cause and impact of vulnerabilities. In their ontology vulnerability management (OVM), they have populated all vulnerabilities of NVD (National Vulnerability Database), with additional inference rules, knowledge representation, and data-mining mechanisms. Another relevant work in vulnerability classification area is done by [10], they proposed a CVE categorization framework that transforms the vulnerability dictionary into a classifier that categorizes CVE(Common Vulnerability and Exposure) with respect to diverse taxonomic features and evaluates general trends in the evolution of vulnerabilities. [11], in their paper, entitled "Secure software Design in Practice" presented a SODA (*a Security-Oriented Software Development Framework*), which was the result of a research project where the main goal had been to create a system of practical techniques and tools for creating secure software with a special focus on the design phase of the software. Another approach of categorizing vulnerabilities is that of [12]. In their paper [12], they looked at the possibilities of categorizing vulnerabilities in the CVE using SOM. They presented a way to categorize the vulnerabilities in the CVE repository and proposed a solution for standardization of the vulnerability categories using a data-clustering algorithm. [13], proposed SecureSync, an automatic approach to detect and provide suggested resolutions for recurring software vulnerabilities on multiple systems sharing/using similar code or API libraries. There are many other vulnerability classification approaches like [14,15,16], but all the above mentioned approaches are either to generic in nature or they cannot be used to classify vulnerabilities on the basis of security properties of the software. Therefore in this research work we are proposing a classifier that is developed using machine learning techniques and is very specific to the design phase of the software. In the next section, a development stage of the classifier is explained.

3.0 DESIGN LEVEL VULNERABILITIES CLASSIFICATION MODEL

Software vulnerability databases are essential part of software security knowledgebase. There are a number of vulnerability databases that exchanges software vulnerability information however, its impact on vulnerability analysis is hindered by its lack of categorization and generalization functionalities [10]. To extract useful and relevant information from these databases, lots of manual work is required. For example, if software developer wants to know about the most common and severe vulnerability prevailing in a current software in a particular period of time then he has to study all the vulnerability descriptions published during that period, then he has to classify those vulnerability based on his own criteria and then he has to check the severity rating provided by various experts. This is very unreliable, tedious and protracted task. Using a proposed classification model, researchers and developers can easily classify design level vulnerabilities and identify a mitigation mechanism in the form of design pattern, in the early phases of the SDLC. The classification results with severity rating can further be used to calculate the risk of vulnerability occurrence at the design phase of the software.

Automated text classifier is basically used to classify the text in the predefined classes. An abstract view of the classifier is shown in Fig 3.0 (a). In proposed design level vulnerability classifier, first text is pre-processed using various processes like tokenization, case transformation, stop-word removal and stemming, then SVM (Support Vector Machine) is used to classify the text and finally bootstrap validation is used to test and validate the results. The development process of the classifier is explained in Fig 3.0 (b).

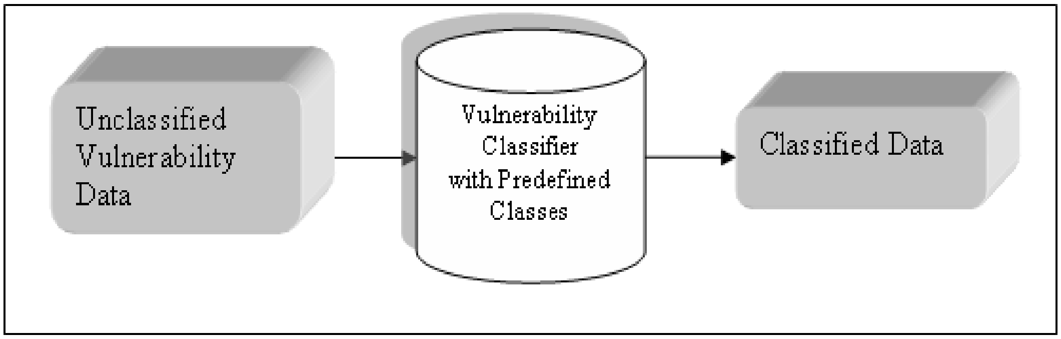

Unclassified Vulnerability Data
Vulnerability Classifier with Predefined Classes
Classified Data

FIGURE 3.0 (a): Abstract Vulnerability Classifier

The vulnerability categorization framework proposed by [10] is similar to this design level classifier. But Chen's framework is a generalized categorization framework that is developed to classify all the vulnerabilities of CVE, on the bases of classification categories of BID, X-force and Secunia. The training data in their framework is also taken from these vulnerabilities databases only.

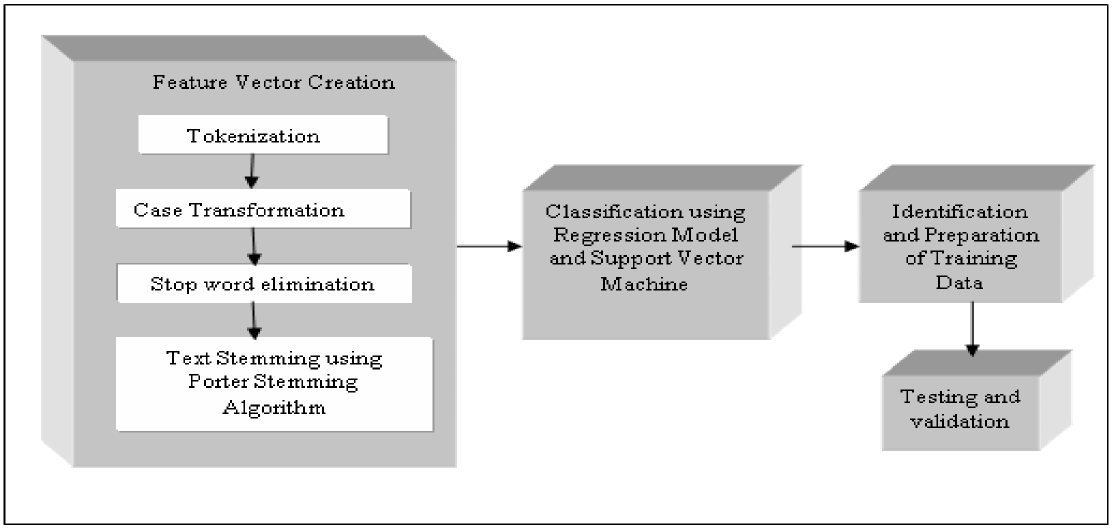

Feature Vector Creation
Tokenization
Case Transformation
Stop word elimination
Text Stemming using Porter Stemming Algorithm
Classification using Regression Model and Support Vector Machine
Identification and Preparation of Training Data
Testing and validation

FIGURE 3.0 (b): Design Level Vulnerabilities Classification Process

In our design level vulnerability classifier, only design level vulnerabilities are classified and in training data only those identified vulnerabilities are considered which can be mitigated at the design level of the software. Moreover the classes are defined on the basis of security properties of the software like authentication, authorization etc., which are generally considered while developing the security design patterns of the software. Therefore after classification developers/researchers can priorities prevailing vulnerabilities class before choosing security design pattern.

3.1 Feature Vector Creation

The vulnerabilities in the CVE are defined in the natural language form. Therefore only way to identify a feature vector using the vulnerability description is the frequency of keywords in the description. Therefore feature vector are identified by the keywords used in the description of the vulnerabilities. To make vulnerability description into a structured representation that can be used by machine learning algorithms, first the text will be converted into tokens and after stemming and stop word removal, and case transformation. There are five steps in the feature creation process, specified as follows:

a). Tokenization
b). Case Transformation
c). Stopword elimination
d). Text stemming of CVE entries
e). Weight Assignment

a).Tokenization

The isolation of word-like units form a text is called tokenization. It is a process in which text stream is to break down into words, phrases and symbols called tokens [17]. These tokens can be further used as input for the information processing. In order to convert text in machine learning form, first the raw text is transformed into a machine readable form, and first step towards it is a tokenization. As shown in Fig. 3.1 (a), the raw text is first feed to the pre-processor, convert the text in the form of tokens then further morphological analysers are used to perform required linguistic analysis.

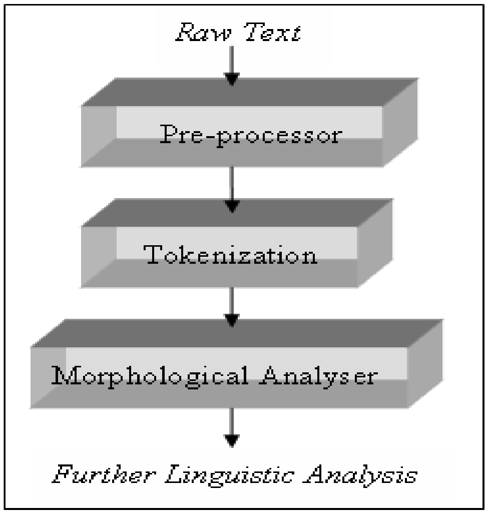


FIGURE 3.1 (a): Text transformations before linguistic analysis

In order to feed vulnerability description in machine learning process, the textual description of vulnerability is first converted in the form of tokens. In Table 3.1 (a), a vulnerabilities description is shown after tokenization.

Vulnerability ID	Vulnerability Description	Vulnerability Description after Tokenization
CVE-2007-0164	Camouflage 1.2.1 embeds password information in the carrier file, which allows remote attackers to bypass authentication requirements and decrypt embedded steganography by replacing certain bytes of the JPEG image with alternate password information.	Camouflage, embeds, password, information, in, the, carrier, file, which, allows, remote, attackers, to, bypass, authentication, requirements, and, decrypt, embedded, steganography, by, replacing, certain, bytes, of, the, JPEG, image, with, alternate, password, information.

TABLE 3.1 (a): Tokenization of Vulnerabilities

b). Case Transformation

When raw text is retrieved for processing from any source, then it contains words in both upper case as well as lower case. The machine learning algorithms reads words in different cases as different words. In order to transform all the words in the same case, Case transformation process is used. Case transformer, transforms all characters in a document to either lower case or upper case, respectively. In our case we have transformed all the words in the document in lower case.

Vulnerability ID	**Vulnerability Description**	**Vulnerability Description after Case Transformation**
CVE-2007-0164	Camouflage, embeds, password, information, in, the, carrier, file, which, allows, remote, attackers, to, bypass, authentication, requirements, and, decrypt, embedded, steganography, by, replacing, certain, bytes, of, the, JPEG, image, with, alternate, password, information	camouflage, embeds, password, information, in, the, carrier, file, which, allows, remote, attackers, to, bypass, authentication, requirements, and, decrypt, embedded, steganography, by, replacing, certain, bytes, of, the, jpge, image, with, alternate, password, information

TABLE 3.1 (b): Case Transformation of Vulnerabilities Description

c). Stop word Removal

Stop word elimination is a process of removing those tokens that are considered as only for grammatical function without adding new meaning to sentences they involve [18]. The stop word list generally consists of articles, case particles, conjunctions, pronouns, auxiliary verbs and common prepositions. There is no unique list of stop words which is always used. There are number of lists that are proposed by different researchers. A list of 418 stop word is used by Chen [10]. A similar stop word list is used in information retrieval systems Snowball [19] and Lemur [20].The vulnerability description after stop word removal is shown in Table 3.1 (c)

Vulnerability ID	Vulnerability description after tokenization and case transformation	Vulnerability description after Stopword Removal
CVE-2007-0164	camouflage, embeds, password, information, in, the, carrier, file, which, allows, remote, attackers, to, bypass, authentication, requirements, and, decrypt, embedded, steganography, by, replacing, certain, bytes, of, the, jpge, image, with, alternate, password, information.	camouflage, embeds, password ,information, carrier, file, allows, remote, attackers, bypass, authentication, requirements, decrypt, embedded, steganography, replacing, bytes, jpge, image, alternate, password, information.

TABLE 3.1 (c): Stop word removal from vulnerabilities

d). Text Stemming

Uses of stemming algorithms in modern information retrieval (IR) systems are common these days. Stemming algorithms are helpful for free text retrieval, where search terms can occur in various different forms in the document collection. Stemming makes retrieval of such documents independent from the specific word form used in the query [21]. To extract the information from vulnerability description, stemming algorithm can be used, so that text can be easily transformed into a machine readable form. Porter stemming algorithm is one of the popular algorithms that is generally used in the information retrieval process. Porter's algorithm consists of 5 phases of word reductions, applied sequentially. Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix. In the first phase, this convention is used with the following rule group [22]:

Rule	Example

SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

The vulnerability description after applying porter stemming algorithm is shown in Table 3.1 (d).

Vulnerability ID	Vulnerability Description	Vulnerability Description after Text Stemming
CVE-2007-0164	Camouflage 1.2.1 embeds password information in the carrier file, which allows remote attackers to bypass authentication requirements and decrypt embedded steganography by replacing certain bytes of the JPEG image with alternate password information.	camouflag emb password inform carrier file allow remot attack bypass authent requir decrypt embed steganographi replac byte jpeg imag altern password inform

TABLE 3.1 (d): Stemming of words in Vulnerabilities Description

e). Weight Assignment

After text stemming, the next step is a weight assignment of each word of vulnerability description. The simplest approach is to assign the weight to be equal to the number of occurrences of term '*t*' in document *'d'*. This weighting scheme is referred to as *term frequency* and is denoted '*tf* $_{t,d}$' with the subscripts denoting the term and the document in order [22]. It is normally computed as follows.

$$tf_{t,d} = f_{t,d} \,/\, max_k\, f_{k,d} \qquad \text{Eq. 3.1 (e.1)}$$

where

$f_{t,d}$ is the frequency (number of occurrence of the term '*t*' in document '*d*') and

$max_k\, f_{k,d}$ (maximum number of occurrences of any term)

Thus, the most frequent term in document 'd' gets a TF as 1, and other terms get fractions as their term frequency for this document. But the disadvantage of using a term frequency is that, in this all the terms are considered equally important. In order to avoid this bias, term frequency and inverse document frequency *(tf-idf)* weighting is used. The IDF for a term is defined as follows [23].

$$Idf_{t,d} = \frac{\log |\, D\,|}{|\ \{d: t \in d\}\ |} \qquad \text{Eq. 3.1 (e.2)}$$

where

- $|\, D\,|$: the total number of documents
- $|\{d : t \in d\}|$: number of documents where the term *t* appears (i.e., $\mathrm{tf}(t,d) \neq 0$)

As defined in [22], the *tf-idf* weighting scheme assigns the term '*t*' a weight in document 'd' given by

$$tf\text{-}idf_{t,d} = tf_{t,d} \,.\, idf_t$$

In other words, term frequency-inverse term frequency assigns to term *'t'* a weight in document *'d'* is

- high when term occurs many times within a small number of documents ;

- lower when the term occurs fewer times in a document, or occurs in many documents;
- lowest when the term occurs in virtually all the documents.

Words Row No.	abil	absenc	Accept	Access	accoria	account	action	…
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…
3	0.0	0.0	0.0	0.06943462298821593	0.0	0.0	0.0	…
4	0.0	0.0	0.0	0.1208492763637866	0.0	0.0	0.0	…
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…
7	0.0	0.0	0.0	0.0711442566706304	0.0	0.0	0.0	…
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…
9	0.0	0.0	0.0	0.12993237910966035	0.0	0.0	0.0	…
10	0.0	0.0	0.0	0.07918346369069837	0.0	0.2787299435038924	0.0	…
…	…	…	…	…	…	…	…	…

TABLE 3.1 (e): Example set of feature vector with their tf-idf

Now after implementing all the above processing we get each vulnerability description as a *vector*, weight that is calculated on using tf-idf formula. Example set of ten rows is shown in table 3.1 (e). This vector form will be used in the scoring and ranking of vulnerabilities.

3.2 Categorization Using Support Vector Machine

Text categorization is the process of categorizing text documents into one or more predefined categories or classes. Differences in the results of such categorization arise from the feature set chosen to base the association of a given document with a given category [24]. There are a number of statistical classification methods that can be applied to text categorization, such as Naïve Bayesian [25], Bayesian Network [25], Decision Tree [26, 27], Neural Network [28], Linear Regression [29], k-NN [30]. SVM (support vector machine) learning method introduced by [31], are well-founded in terms of computational science. Support vector machines have met with significant success in numerous real-world learning tasks [25]. Compared with alternative machine learning methods including Naive Bayes and neural networks, SVMs achieve significantly better performance in terms of generalization [32, 33]

SVM classification algorithms, proposed by Vapnik [34] to solve two-class problems, are based on finding a separation between hyperplanes defined by classes of data, shown in Figure 3.2 (a).

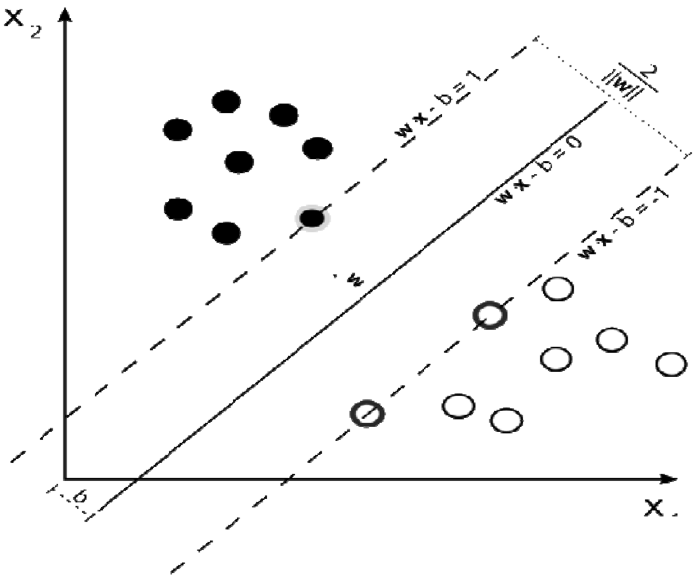


FIGURE 6.2.2: Example of SVM hyper plane pattern

This means that the SVM algorithm can operate even in fairly large feature sets as the goal is to measure the margin of separation of the data rather than matches on features [24]. The SVM is trained using pre-classified documents.

As explained in [34], for a given set of training data $T = \{.xi, yi\}$ ($i = 1,\ldots, m$), each data point $.xi$ Є Rd with d features and a true label yi Є $Y = \{l1, \ldots, lk\}$. In case of binary classifier, label set is $Y = \{l1 = -1, l2 = +1\}$, it classifies data point in positive and negative by finding separating hyperplane. The separating hyperplane can be expressed as shown in Eq. 3.2 (a).

$\overline{w} \cdot .\overline{x} + b = 0$ --Eq 3.2(a)

where w Є R^d *is a* weight vector is normal to the hyperplane, operator (·) computes the inner-product of vectors w and x and b is the bias. Now we want to choose the '*w*'and *'b'* to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible while still separating the data. These hyperplanes can be described by the equations

$w \cdot .x - b = 1$--Eq 3.2 (b)

and

$w \cdot .x - b = -1$---Eq 3.2 (c)

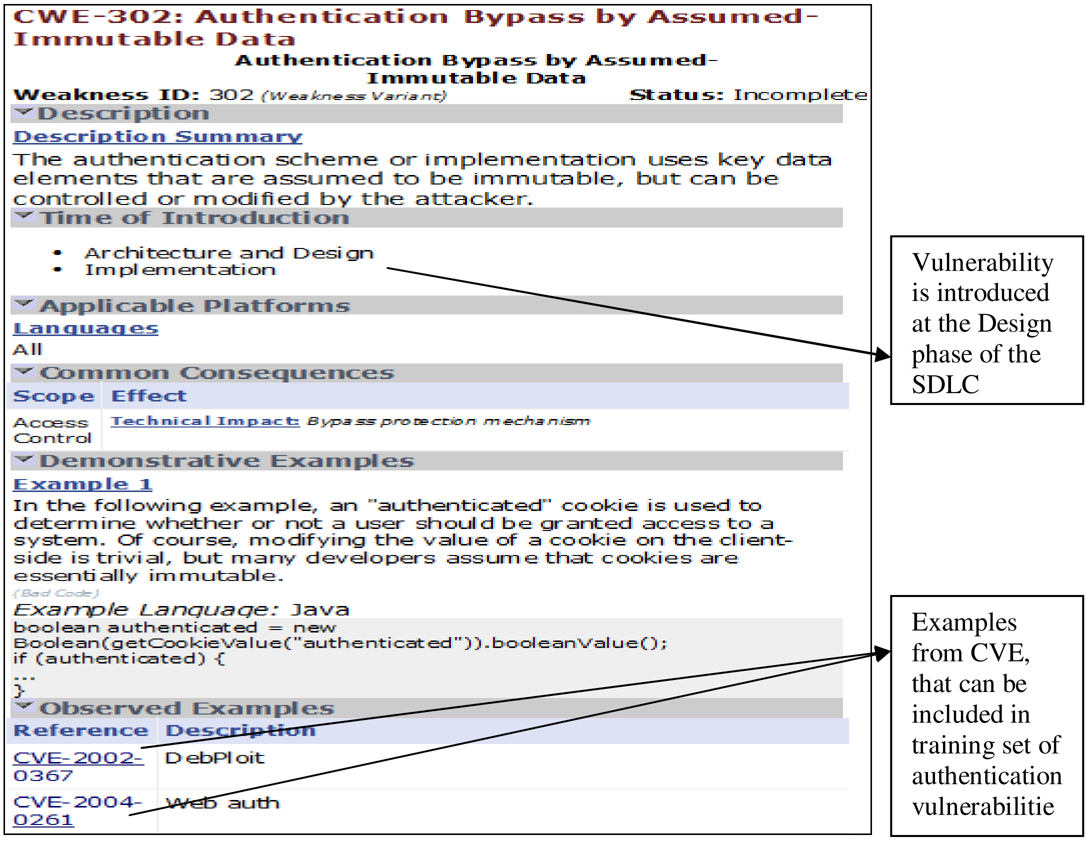
In the case that the label set $Y = \{l1 = 1, \ldots, lk = k\}$ and $k > 2$, a multiclass SVM learning model is built with two approaches in the proposed framework: *multiclass-to binary reduction* and *multiclass-optimization* methods [10]. In the multiclass-to-binary reduction method, the learning problem in question is reduced to a set of binary classification tasks and a binary classifier is built independently for each label *lk* with the one-against-rest training technique [35]. When more than two classes are involved then regression model is used. regression model builds a classifier using a regression method which is specified by the inner operator. For each class *i* a regression model is trained after setting the label to (+1) if the label equals *i* and to (-1) if it is not. Then the regression models are combined into a classification model. Here we are using SVM classification model, therefore Regression model is combined into SVM. In order to determine the prediction for an unlabeled example, all models are applied and the class belonging to the regression model which predicts the greatest value is chosen.

3.3 Identification and Preparation of Training Data

There are number of public and private vulnerability databases that classify vulnerabilities on different bases like cause, phase of occurrence, product specific etc. the list is shown in Table 3.3 (a). But they all are too generic in nature. The CWE (Common weakness Enumeration) [36] is a vulnerability database and portal in which each vulnerability is specified with its type, mitigation, phase of introduction and security property it belongs to. Therefore it is a best place from where vulnerability data can be collected on the basis of phase of introduction and the security property it belongs to. Fig 3.3 (a) is screenshot of the CWE window, it is showing a information that each entry of CWE contains. In this we are interested in only those vulnerabilities that can be mitigated in the design phase of the software. But in CWE the vulnerabilities are divided into number of classes and number of examples are given in the description of each class. In order to collect the training data from CWE, we explore the required security class, then collect vulnerability example from each class. Almost all the examples that are used in CWE are from CVE (Common Vulnerability and Exposure).Maximum possible numbers of examples are collected from the CWE for the training set.

S. No.	Database Name	URL
1.	Common Vulnerability and Exposures	http://cve.mitre.org/
2.	Common Weakness Enumeration	http://cwe.mitre.org/
3.	Computer Associates Vulnerability Encyclopedia	http://www3.ca.com/securityadvisor/vulninfo/browse.aspx
4.	Dragonsoft vulnerability database	http://vdb.dragonsoft.com
5.	ISS X-Force	http://xforce.iss.net/xforce/search.php
6.	National Vulnerability Database	http://nvd.nist.gov/
7.	Open Source Vulnerability Database	http://www.osvdb.org/
8.	Public Cooperative Vulnerability Database	https://cirdb.cerias.purdue.edu/coopvdb/public/
9.	Security Focus	http://www.securityfocus.com/vulnerabilities/

TABLE 3.3 (a): Vulnerability Database with their URLs



CWE-302: Authentication Bypass by Assumed-Immutable Data

Authentication Bypass by Assumed-Immutable Data

Weakness ID: 302 *(Weakness Variant)* Status: Incomplete

Description

Description Summary

The authentication scheme or implementation uses key data elements that are assumed to be immutable, but can be controlled or modified by the attacker.

Time of Introduction

- Architecture and Design
- Implementation

Applicable Platforms

Languages

All

Common Consequences

Scope	Effect
Access Control	Technical Impact: *Bypass protection mechanism*

Demonstrative Examples

Example 1

In the following example, an "authenticated" cookie is used to determine whether or not a user should be granted access to a system. Of course, modifying the value of a cookie on the client-side is trivial, but many developers assume that cookies are essentially immutable.

(Bad Code)

Example Language: Java

```
boolean authenticated = new
Boolean(getCookieValue("authenticated")).booleanValue();
if (authenticated) {
...
}
```

Observed Examples

Reference	Description
CVE-2002-0367	DebPloit
CVE-2004-0261	Web auth

Vulnerability is introduced at the Design phase of the SDLC

Examples from CVE, that can be included in training set of authentication vulnerabilitie

FIGURE 3.3 (a): CWE Vulnerability Class Description Window

After the exhaustive search of CWE vulnerability classes, the number of vulnerability examples that are collected, are shown in Table 3.3 (b). While collecting data from CWE, at most care is taken to include only those vulnerability classes where the time of introduction is specified as "design phase"

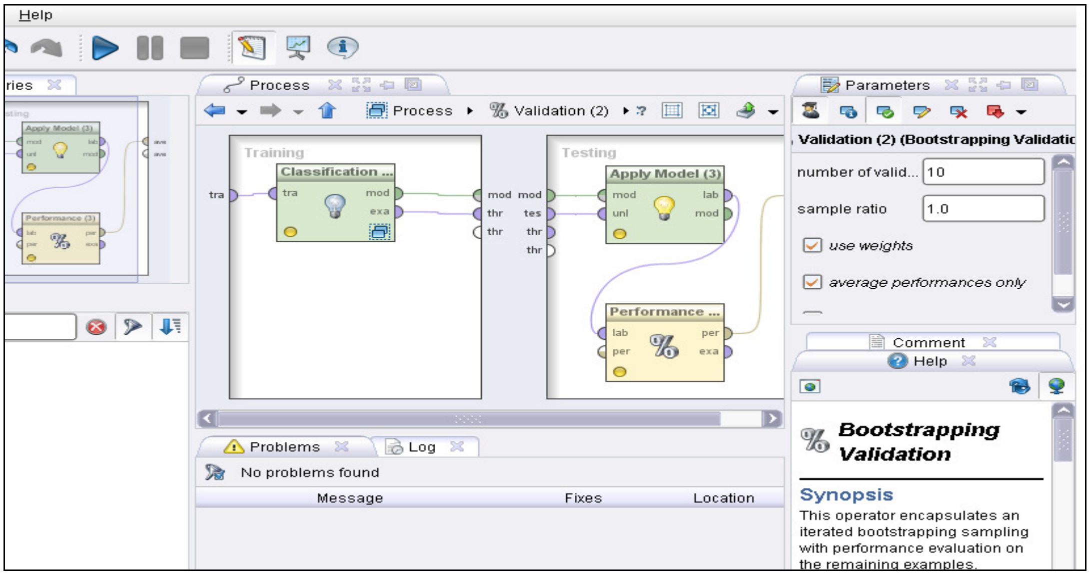
S. No.	Vulnerability Class	Number of Training data
1.	Authentication	54
2.	Authorization	50
3.	Audit and Logging	36
4.	Secure Information Flow	31
5.	Secure Session Management	24

TABLE 3.3 (b): Number of training data identified under each class

3.4 Testing and Validation

After the collection of training data, a SVM learning model can be built. But SVM is basically a binary classifier. As we have multiple classes with $|Y|>2$, the learning

problem is decomposed into $|Y|$ binary classification tasks with the multiclass-to-binary reduction method, and subsequently $|Y|$ binary classifiers are built with the one-against-rest training method that essentially transforms the learning task for category l_i of Y into a two-class categorization by treating data points with label l_i in the training data as positive examples and the remaining data points as negative examples [10]. We are using Rapidminer tool to implement the SVM. Regression model is used to classify the data into multiple class. After supplying the data regression model, bootstrap validation is used to validate the classification. Fig 6.2.4 (a) is showing the screen shot of rapid miner while implementing bootstrap validation.



Validation (2) (Bootstrapping Validatic

number of valid... 10

sample ratio 1.0

☑ use weights

☑ average performances only

No problems found

Message | Fixes | Location

Bootstrapping Validation

Synopsis

This operator encapsulates an iterated bootstrapping sampling with performance evaluation on the remaining examples.

FIGURE 3.4 (a): Screen shot from 'Rapidminer Tool', while implementing bootstrap validation

3.5 Bootstrap Validation

The bootstrap family was introduced by Efron and is first described in [37]. In this method, for given dataset of size n a bootstrap sample is created by sampling n instances uniformly from the data. There are several bootstrap methods. A commonly used one is the 0.632 bootstrap. As explained in by Han and Kimber in their book 'Data Mining: Concepts and Techniques' [38], in this method, suppose we have 'd' tuples. The data set is sampled 'd'/6 times, with replacement, resulting in the bootstrap samples or the training set of d samples. The data tuple that are not included in the training, forms the test set. Now the probability for each tuple to be selected is 1/d, and the probability of not being chosen is (1- 1/d).We have to select d times, so the probability that a tuple will not be chosen during this whole time is (1-1/d)d. If d is large, the probability approaches e^{-1} = 0.368. Thus, 36.8% of tuples will not be selected for training and thereby ends up in the test set, and the remaining 63.2% will form the training set.

The sampling procedure can be recorded k times, where in each iteration, we can use the current test set to obtain the accuracy estimate of the model obtained from the current bootstrap sample. The overall accuracy of the model is then estimated as

$$\text{Acc (M)} = \sum_{i=1}^{k} (0.632 \text{ X Acc } (M_i)_{test\text{-}set} + 0.368 \text{ X Acc}(M_i)_{train\_set}) \qquad \text{Eq. 3.5 (a)}$$

where Acc $(M_i)_{test\text{-}set}$ is the accuracy of the model obtained with the bootstrap sample 'i' when it is applied to the test set 'i'. Acc$(M_i)_{train_set}$ is the accuracy of the model obtained with bootstrap sample 'i' when it is applied to the original set of the data tuples. The bootstrap method works well with the small data set.

The whole process that is followed in making the classifier is shown in Figure 3.5 (a). the rapid miner data mining tool is used that have almost all the available data-mining process in the form of operators. First of all training data is feed to the regression model that is integrated with SVM, then 'Apply Model' operator is used to apply the created model and performance operator is used to measure the performance of the classifier. As an output, the confusion matrix is be obtained that will show the accuracy of the classifier.

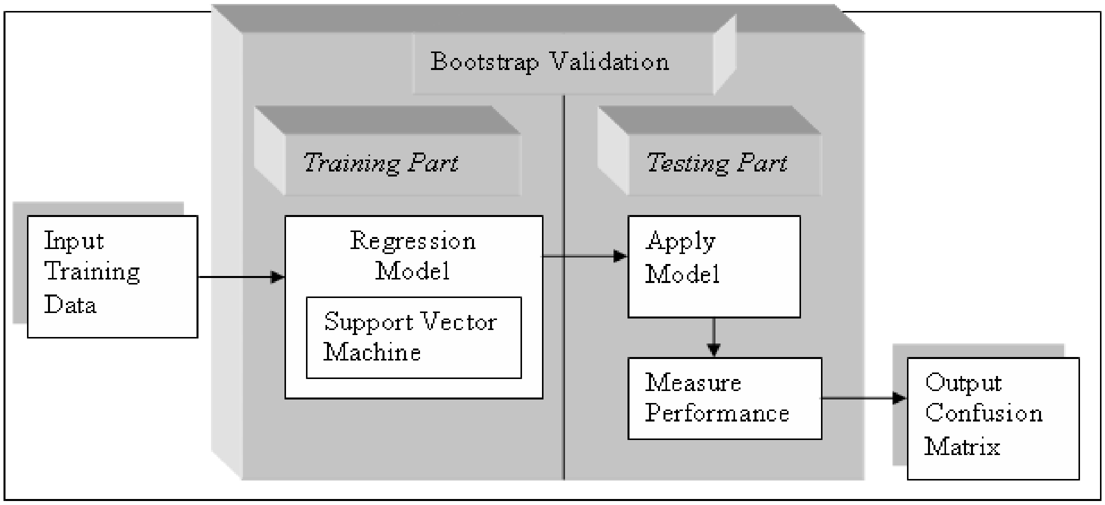


FIGURE 3.5 (a): Classification Model using Bootstrap Validation

The confusion matrix that is obtained after the application of the classifier is shown in Table 3.5 (a) and the 3D- graphical representation of the confusion matrix is shown in Fig 3.5 (b).

True Pred.	True Authorization	True Others	True Secure-Information-Flow	True Audit and Logging	True Authentication	True Session-management	Class Precision
Pred. Authorization	177	2	0	0	4	0	96.72%
Pred. Others	0	139	0	0	2	0	98.58%
Pred. Secure-Information-Flow	0	0	131	0	0	0	100.00%
Pred. Audit and Logging	0	4	0	128	7	0	92.09%
Pred. Authentication	2	16	0	6	189	0	88.73%
Pred. Session-management	0	4	0	0	3	92	92.93%
Class Recall	98.88%	84.24%	100.00%	95.52%	92.20%	100.00%	

Accuracy: 94.52% +/- 1.85% (mikro: 94.48%)

TABLE 3.5 (a): Confusion Matrix

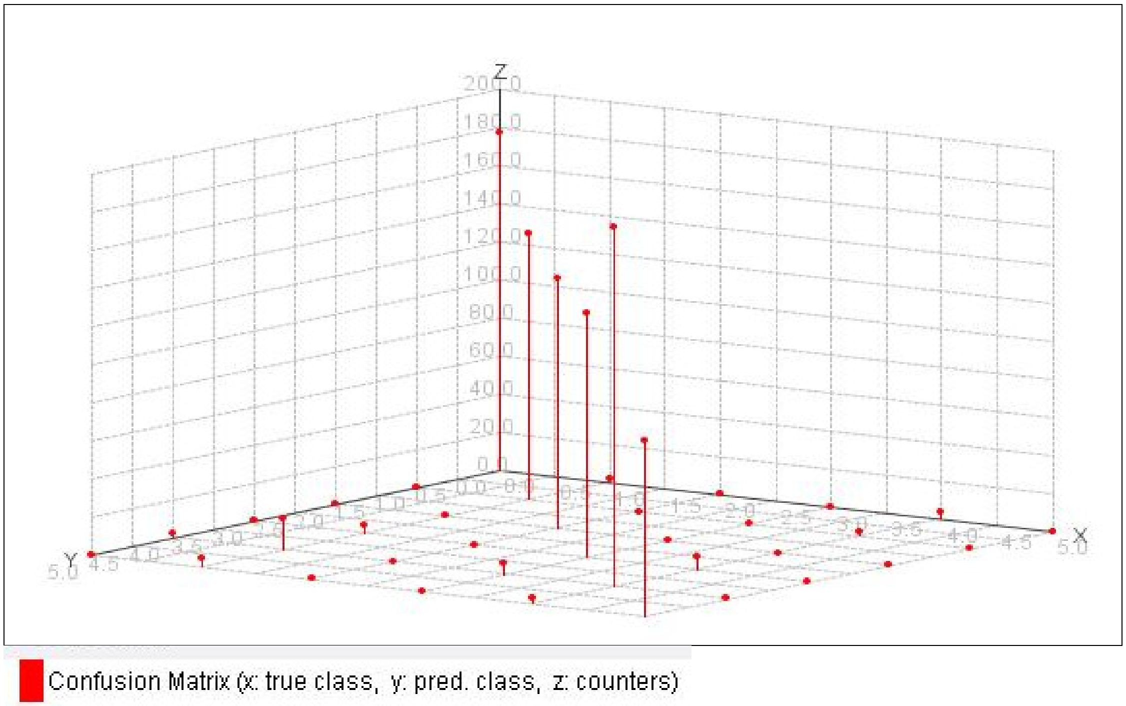


FIGURE. 3.5 (b): Confusion matrix in the form of 3D Graph

All most all the classes have class precision value above 90%. The accuracy rate of about 90% makes the classifier quite accurate (Han and Kamber, 2006). The overall accuracy rate of developed classifier is 94.5 %.

As shown in Table 3.5 (a), the class precision of 'authentication class' is only 88%, because the keywords used in the authentication class are common to other classes also. For example the vulnerability description mainly consist of words like '*unauthenticated user*', '*not allowed authenticated user*', etc, which actually don't indicate the cause as authentication, but classifier gets confused due to the frequent use of theses terms in other classes also, which affect the performance of classifier. But overall accuracy of the classifier is acceptable, which is 94.5%.

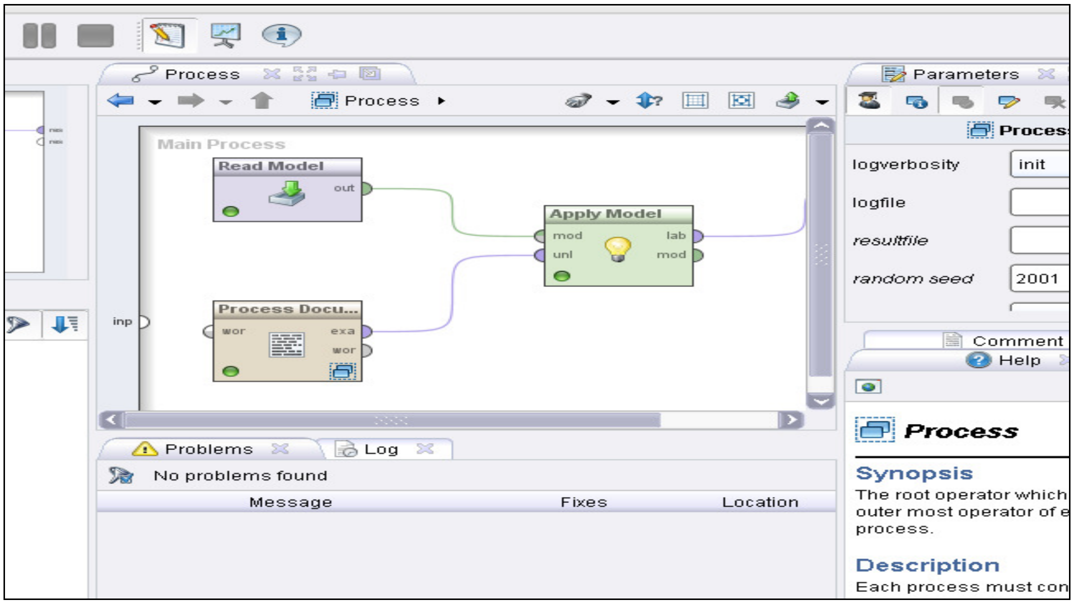
4.0 CLASSIFICATION RESULTS

Now using this design level 'Vulnerability Classification Model' the vulnerabilities can be classified into six classes. In NVD (National Vulnerability Database), total 427 vulnerabilities are identified as design level vulnerabilities till February 2009. Now in order to classify these vulnerabilities in our predefined six classes, vulnerabilities first need to be feed in the classifier, then predicted values can be used for further analysis. After feeding 427 design level vulnerabilities in the model, the example set of the predicted values that is obtained is shown in Table 4.0 (a). The screenshot from Rapidminer during the application of the classifier is shown in Fig.4.0 (a) and the final number of classified vulnerabilities is shown in Fig.4.0 (b). From the classification results it is clear that out of 427 vulnerabilities that are classified as design level vulnerabilities, 117 are actually not design level vulnerabilities. From remaining vulnerabilities, Authentication and authorization related vulnerabilities are most prevailing one, constituting about 53% of total vulnerabilities.

C* V.No.	Confidence (Authorization)	Confidence (Others)	Confidence (Secure Information Flow)	Confidence (Audit and logging)	Confidence (Authentication)	Confidence (Session Management)	Prediction
	-0.703822126777	-0.7411796480	-0.05035211037	1.0	-0.995248931575	0.0	Audit and Logging
2	-0.466536388761	1.0	-1.04654845455	-0.54866717509	-0.733920948135	0.0	Others
3	-0.803923996569	-0.5640368166	-0.61342764113	-0.39066602401	1.0	0.0	Authentication
4	-0.927380058636	-0.8005797124	-0.56610700139	0.01646968988	1.0	0.0	Authentication
5	-0.631178417772	-0.72385838954	-0.99080223010	-0.45735700877	1.0	0.0	Authentication
6	-0.674072540417	-1.0328340541	-1.03255019721	0.32822280439	1.0	0.0	Authentication
7	-1.022489071514	-0.9311227632	-0.13558666871	1.0	-0.563459046881	0.0	Audit and Logging
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…

TABLE 4.0 (a): Sample dataset from classification result

Percentage of Authentication is 30%, which makes it most important property to be mitigated at the design phase of the software. Authorization constitute 23.2% of all the vulnerabilities, which makes it second most important security attribute to be



logverbosity init
logfile
resultfile
random seed 2001

No problems found

Message Fixes Location

Process

Synopsis

The root operator which
outer most operator of e
process.

Description

Each process must con

FIGURE 4.0 (a): Screenshot from rapid miner, while implementing the final model

The percentage of ‘audit and logging’, ‘secure information flow’ and ‘session management’ are 18%, 15% and 12% respectively, which makes them almost equally important.

Vulnerability Class	Count	Percentage	Percentage excluding Others
Authentication	96	22.48244	30.96774
Authorization	72	16.86183	23.22581
Audit and Logging	56	13.11475	18.06452
Secure-Information-Flow	47	11.00703	15.16129
Session-management	39	9.133489	12.58065
Others	117	27.40047	0.0
Total	427	100	100

TABLE 4.0 (b): Number of vulnerabilities classified under each class

These vulnerability classification data can be used with the severity rating to calculate the risk of vulnerability occurrence at the design phase.

5.0 CONCLUSION AND FUTURE WORK

As discussed in the previous sections, study of known vulnerabilities is very useful tool for the developer. Our approach is in the direction of identifying, classifying and learning from known vulnerabilities. So that these vulnerabilities can be avoided in the next generation of the software. In available vulnerability databases, there is no information about the vulnerability cause or the SDLC phase in which they can be removed. Using our proposed classification model, developer would be able to classify any vulnerability from any vulnerability database. The classification model will tell the developer whether the vulnerability be mitigated at the design level? If vulnerability is identified as a design level vulnerability then it will classify the identified vulnerability in security feature. After knowing the class of security feature, designer can adapt necessary design patterns that can prevent these vulnerabilities in the new under-developed software. The accuracy of the classified is found to be satisfactory and it can be used to classify future vulnerabilities. Classifying ‘exposure leading to access violation’ class of vulnerabilities is one of the prompt future works that can be done. The classification results can further be used to calculate the security risk at the design phase of the software. After risk calculation, mitigation mechanisms in the form of design patterns can be identified and thus designer will be able to mitigate security vulnerabilities at the design phase of the software. Another future work that can be done is the creation of tool that can automate the task of vulnerability classifications. After this classification our prompt objective will be to identify, analyze and classify design patterns that can be adapted in order to avoid vulnerabilities in the new software.

REFERENCES

[1] P.T. Devanbu and S. Stubblebine, “Software Engineering for Security: a Roadmap”. International Conference on Software Engineering 2000 special volume on the Future of Software Engineering, 2000, pp.227-239.

[2] G. Hoglund and G. McGraw. “Exploiting Software: How to Break Code”, New York: Addison-Wesley, 2004

[3] L. Lowis and R. Accorsi. “On a Classification Approach for SOA Vulnerabilities”, 33rd Annual IEEE International Computer Software and Applications Conference. 2009, pp 439-444.

[4] V.C. Berghe, J. Riordan and Piessens “A Vulnerability Taxonomy Methodology applied to Web Services”, 10th Nordic Workshop on Secure IT Systems, 2005.

[5] N. Moha. “Detection and Correction of Design Defects in Object-Oriented Designs”. Doctoral Symposium, 21st International Conference on Object-Oriented Programming, Systems, Languages and Application, 2007.

[6] I.V. Krsul, “Software Vulnerability Analysis”. Ph.D. Thesis. Purdue University. USA, 1998.

[7] S. Rehman, and K.Mustafa. “Software Design Level Security Vulnerabilities”, International Journal of Software Engineering, 4 (2). 2011.

[8] T. Joachims. “Text categorization with support vector machines: learning with many relevant features”. 10th European Conference on Machine Learning. 1998.

[9] J. A. Wang, and M. Guo. “OVM: An Ontology for Vulnerability Management”. 7th Annual Cyber Security and Information Intelligence Research Workshop.Tennessee, USA. 2009.

[10] Z. Chen, Y. Zhang, and Z. Chen “A Categorization Framework for Common Computer Vulnerabilities and Exposures”. Computer Journal Advance Access, 2009. Available: http://comjnl.oxfordjournals.org/ cgi/content/abstract/bxp040.

[11] P.H. Meland, and J. Jensen. “Secure Software Design in Practice”. Third International Conference on Availability, Reliability and Security. 2008.

[12] Y. Li, H.S. Venter, and J.H.P Eloff. “Categorizing vulnerabilities using data clustering techniques”, Information and Computer Security Architectures (ICSA) Research Group. 2009.

[13] N.H.Pham, T.T Nguyen, H.A Nguyen,., X.Wang, , A.T. Nguyen, and T.N Nguyen. “Detecting Recurring and Similar Software Vulnerabilities”, International Conference of Software Engineering. Cape Town, South Africa. 2010.

[14] D. Byers, S. Ardi, , N. Shahmehri and C. Duma. “Modelling Software Vulnerabilities with Vulnerability Cause Graphs”. 22nd IEEE International Conference on Software Maintenance. , 2006.

[15] V. Sridharan, and D.R. Kaeli . “Quantifying Software Vulnerability”. Workshop on Radiation effects and fault tolerance in nanometer technologies, Ischia, Italy, 2008.

[16] Y.Wu, R.A. Gandhi, and H. Siy. “Using Semantic Templates to Study Vulnerabilities Recorded in Large Software Repositories”. 6th International workshop on software Engineering for secure system, Cape Town, South Africa. 2010.

[17] G. Grefenstette and P. Tapanainen. “What is a Word, What is a Sentence? Problems of Tokenization”. 3rd Conference on Computational Lexicography and Text Research . 1994, pp. 79-87.

[18] C. Fox. “Lexical Analysis and Stoplist-Data Structures and Algorithms”. New York: Prentice-Hall. 1992.

[19] M. F. Porter. “Snowball: A string processing language for creating stemming algorithms in information retrieval”, 2008. Available: http://snowball.tartarus.org.

[20] Lemur Project (2008). The Lemur Toolkit: For Language Modeling and Information Retrieval, 2008. Available: http://www.lemurproject.org.

[21] M. Braschler and B. Ripplinger, “How Effective is Stemming and Decompounding for German Text Retrieval”. Information Retrieval, 7, 2003, pp.291–316.

[22] C.D. Manning, P. Raghavan, and H. Schütze. “Introduction to Information Retrieval”, Cambridge University Press. 2008.

[23] A. Rajaraman, and J.D. Ullman, Mining of Massive Datasets. 2010. Available: http://infolab.stanford.edu/~ullman/mmds/ch1.pdf

[24] A. Basu, C. Walters, M. Shepherd. “Support vector machines for text categorization”. 36th Annual Hawaii International Conference,2003

[25] T. Joachims. “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization”, 14th International Conference on Machine Learning. 1997.

[26] J.R. Quinlan. “Programs for machine learning”. San Francisco: Morgan Kaufmann Publishers.1993.

[27] S. M. Weiss, C. Apte, F.J. Damerau, D.E. Johnson, F.J. Oles, T., Goetz, T. Hampp. “Maximizing text-mining performance”. IEEE Intelligent Systems Magazine, 1999.

[28] E. Wiener, J. O. Pederson, A.S. Weigend. “A neural network approach to topic spotting”, 4th Annual Symposium on Document Analysis and Information Retrieval. 1995.

[29] Y. Yang and , J.O. Pederson. “A comparative study on feature selection in text categorization”. International Conference on Machine Learning. 1997.

[30] Y. Yang. “An evaluation of statistical approaches to text categorization”. Journal of Information Retrieval. 1 (2). 1999.

[31] V. Vapnik,. “The Nature of Statistical Learning Theory”. Berlin: Springer. 1995.

[32] C. Burges. "A tutorial on support vector machines for pattern recognition”. Data Mining and Knowledge Discovery, 2, 1998, pp. 1-47.

[33] J.T.K. Kwok. “Automated Text Categorization Using Support Vector Machine”. International Conference on Neural Information Processing, 1998.

[34] V. Vapnik. “Statistical Learning Theory”. New York: John Wiley and Sons. 1998.

[35] T. Hastie, and R. Tibshirani, “Classification by pair wise coupling. Ann. Statist”, 26, 1998, pp. 451–471.

[36] CWE (Common Weakness Enumeration). Available: http://cwe.mitre.org/

[37] B. Efron. “ Estimating the error rate of a prediction rule: Improvement on cross-validation”. Journal of the American Statistical Association, 78, 1983. pp.316-331.

[38] J. Han, and M. Kamber “Data Mining: Concepts and Techniques”. San Francisco: Morgan Kaufmann Publisher, 2006.

Performance Review of Zero Copy Techniques

Jia Song *song3202@vandals.uidaho.edu*
Center for Secure and Dependable Systems
University of Idaho
Moscow, ID 83844-1010 USA

Jim Alves-Foss *jimaf@uidaho.edu*
Center for Secure and Dependable Systems
University of Idaho
Moscow, ID 83844-1010 USA

Abstract

E-government and corporate servers will require higher performance and security as usage increases. Zero copy refers to a collection of techniques which reduce the number of copies of blocks of data in order to make data transfer more efficient. By avoiding redundant data copies, the consumption of memory and CPU resources are reduced, thereby improving performance of the server. To eliminate costly data copies between user space and kernel space or between two buffers in the kernel space, various schemes are used, such as memory remapping, shared buffers, and hardware support. However, the advantages are sometimes overestimated and new security issues arise. This paper describes different approaches to implementing zero copy and evaluates these methods for their performance and security considerations, to help when evaluating these techniques for use in e-government applications.

Keywords: Zero Copy, Network Security, Security/Performance Tradeoffs

1. INTRODUCTION

In addition to growth of corporate servers, there has been tremendous interest and growth in the support of e-government services. This includes remote access to public information, databases, forms, guidance, training materials as well as access to personal information such as treasury holdings, social security, income taxes, and government benefits. In addition, e-government also supports internal access to government documents, support of paperless offices and workflow improvement and communication support for critical services including first responders in emergencies. Cost-effective e-government solutions will require use of shared organizational servers, such as cloud servers, and high performing computers that can provide users with timely and secure access to information. Unfortunately, we have found that there is often a disconnect between marketed performance-improving solutions and security needs. This paper addresses one “performance-improving” technology with respect to security needs. We have found that some of the solutions that have been proposed and implemented will not work with security technologies, such as encryption. The understanding of the impact of performance-improving solutions on security is important when comparing vendors' performance benchmarks and claims when evaluating systems for purchase, and is important for developers to understand when making implementation decisions.

Therefore, the intent of this paper is to bring to light an example of security concerns that we believe should be addressed in the development process and in security requirements such as those specified by US Federal Acquisition Regulation (FAR) Part 39.101 and OMB Circular A-130 while supporting e-government efforts such as those specified in the “E-Government Act of 2002”. Similar regulations exist in other countries or in corporate policies. Corporations and governments in the market for servers should pay careful attention to performance benchmarks developed by vendors.

For example, the FAR Part 39.101 requires that agencies identify their requirements pursuant to: best management practices for energy-efficient management of servers and Federal data centers and shall include the appropriate information technology security policies and requirements. Energy-efficient management often translates into performance, since a data-center can use a lesser number higher performing servers to do the same work, and with less energy. As we discuss in this paper, we need to be aware of the tradeoff between some performance technologies and security technologies. We believe there is a need for the development of a set of benchmarks or at least checklists/configuration guides that enable the acquisition officers and developers to make the most appropriate choice of technologies given the security needs of the application environment.

1.1 Performance and Security

As demand for higher performance computers grew, CPU designs developed from single core to dual-core, quad-core or even hexa-core. However, a quad-core processor does not indicate that the performance of the overall system is four times the performance of a single core processor. The performance of the system is limited by many other factors, such as access to system memory. According to Brose [1], CPU processing power improves about 50% each year, while the memory only has an average 35% increment at the same amount of time[1]. Improvement in memory usage is increasingly a critical factor for system performance. This is especially important for network applications, such as severs, which transfer a lot of data, and cannot benefit from cache size increases.

Servers used by industry and government are critical not only for their performance, but also their security. To make the government more transparent, web servers for the public are used to inform the public about new policies and the latest news from the government. What's more, the servers used among governmental employees must be secure enough to avoid leaking information and to prevent hacking. Therefore, severs used by government must ensure their efficiency and security.

Traditionally, in server environments, such as servers used in e-government, when sending data to a network, data is loaded from disk and transferred to a network card buffer. The whole procedure consumes a lot of CPU cycles and memory bandwidth, for data must be copied between application memory space and kernel memory space. However, most of these data copies are redundant and can be minimized by implementing zero copy techniques. Zero copy is a name used to refer to a variety of techniques which help reduce useless memory accesses, usually involving elimination of data copying. By implementing zero copy, less data copies are needed when data is transferred between buffers. This helps reduce the number of unnecessary memory accesses and CPU consumption, and therefore enhance overall system performance.

1.2 Zero copy for Performance

In recent years, there have been a number of studies regarding zero copy and different kinds of zero copy techniques have been applied. Zero copy can be classified into two groups: (1) reduce data copies between devices (disk or network card) and kernel buffers, and (2) reduce data copies between kernel buffers and application buffers.

Typically, data copies between devices and kernel buffers can be reduced by using DMA (Direct Memory Access). DMA allows special purpose hardware to read or write main memory without involving the CPU, which greatly reduces CPU consumption and also eliminates redundant data copies between device and kernel buffers.

Zero copy can also reduce the number of copies between user memory space and kernel memory space. Several different approaches have been proposed to solve this problem, and they are all called zero copy. For example, memory mapping is used to map the user buffer to the

[1] This means that relative to CPU speeds, memory performance is cut in half every 7 years.

kernel buffer; in LyraNET, a modified `sk_buff` structure can be used to eliminate data copies; and buffers can be shared, so stored data can be referenced via pointers. In addition, some system calls in Linux support zero copy by reducing data copies such as the `sendfile()` and `splice()` system calls. Instead of copying data, descriptors are appended to buffers in sendfile. The system call `splice()` sets up a data transfer path between different buffers in kernel memory.

However, in addition to performance benefits of zero copy techniques, it is important to understand the security implications of using those techniques. This is important for commercial servers as well as e-government systems.

Zero copy helps reduce the number of data copies and context switches which improves the CPU performance. But it also causes some new problems that we have to understand. This paper focuses on analyzing different zero copy techniques which are used to reduce the data copies in TCP/IP transmission, which are being implemented in many commercial servers which could end up being used by government offices to gain better performance.

The remainder of this paper evaluates the use of zero copy for performance in secure environments as is laid out as follows. The traditional data transfer approach and security issues are presented in Section 2 and zero copy techniques are described in Section 3. Problems caused by zero copy are examined in Section 4; this section also discusses security issues relating to zero copy techniques, which will help developers and acquisition offices in determining the appropriate use of this technology.

2. BACKGROUND

2.1 Traditional Data Transfer Method

Traditionally, if a user/server application wants to send data to a network, the data is read from disk to kernel memory, then stored in a buffer allocated by the user application in user memory, then sent to the kernel buffer associated with the network stack, and finally sent to the network interface card (NIC) for transfer over the network. Although the whole procedure needs only two system calls -- `read()` and `write()`, from the view of kernel, it is quite inefficient, even if DMA is supported by hardware. The traditional data transfer method is shown in Figure 1. We will examine this approach step by step to illustrate the CPU copies[2] and context switches.

First, if the data is not cached in a kernel buffer, the kernel needs to load the data from disk to a kernel buffer. For this step, DMA is typically used, which allows this first copy to consume little CPU resources. However, extra effort is needed to manage the DMA copy. Then, the data needs to be transferred into the buffer allocated by the application in user memory. This step is completed by the `read()` call, which, is typically done by a CPU copy in a kernel function such as `copyout()` or `copy_to_user()` [1]. Also, the `read()` system call needs a context switch from user mode to kernel mode. After copying data to the application buffer, the `read()` system call returns, which causes another context switch from kernel mode to user mode.

From the application buffer to the corresponding network stack in the kernel buffer the data needs to be transferred using a user-to-kernel copy function, such as `copyin()` or `copy_from_user()` [1]. This kernel buffer is not the same buffer used for the initial fetch from disk. It is a new buffer which is associated with the network stack. Data is packetized according to the Internet protocol in the network stack. For this step, the use of the `write()` system call generates a context switch from user mode to kernel mode. Finally, the prepared data are sent from the kernel buffer to the buffer on the network interface card, and then transferred over the network. This fourth copy can be done using DMA. Although no CPU copy is needed, the returned `write()`system call forces another context switch from kernel back to user mode.

[2] We use the term CPU copy to refer to any time the CPU reads a memory cell and writes a result back out.

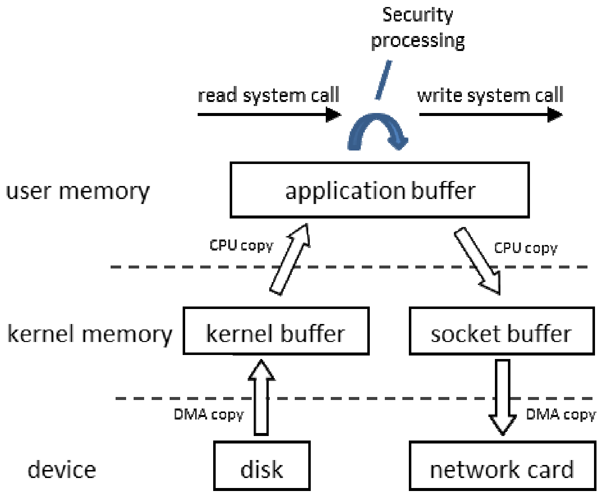


FIGURE 1: Traditional Data Transfer

2.2 Secure Communications

For secure communication across the network, applications will encrypt data packets. Depending upon the security mechanism used, encryption could occur at a high level or low level of the network stack. Servers and secure network applications often use SSL (Secure Socket Layer) for secure communication.

There are two aspects to SSL that can slow down system processing. The first is the negotiation of the encryption key and authentication. These processes typically use modular arithmetic over very large numbers and take a good amount of CPU time. There are hardware accelerators, such as SSL cards, that allow an application to offload this computation load. This set up of the encryption is beyond the scope of this paper. The second aspect is the actual encryption of the data. This requires applying an encryption algorithm to all of the data, reading the results from one buffer and writing them to another buffer (labeled *security processing* in **Error! Reference source not found.**). Even with CPU help, data copies are needed, for example in 2010 Intel introduced the AES instruction set extension to support fast encryption, and utilize 128-bit registers in the Intel architecture. These instructions require CPU access to the data and hence require a CPU copy.

In addition to encrypted communication, some security system requires filtering or guarding by the application. Typically this occurs in a secure communications environment where messages are being transmitted between different security domains. The concepts of zero copy still work, but the source is often another network device and not a disk. Filtering also occurs when data retrieved from a database is filtered before release. In these situations, the application needs to read and review the information before making the decision whether to allow the information to go out (labeled *security processing* in **Error! Reference source not found.**). We also have to make sure that no unauthorized processes get access to the sensitive information. As we examine different zero copy techniques, we will have to see if they provide any performance enhancement in encryption or filtering secure environments.

3. ZERO COPY TECHNIQUES

There are at least four data copies and four context switches during the data transmission in the traditional data transfer approach. According to Huang et al [2], in addition to the costs of data copies, software interrupts consume a great portion of CPU usage, for interrupts are generated for each packet. As we see, some of the data copies are redundant and some of the context switches between user mode and kernel mode can be eliminated by avoiding calling the read and write system calls separately. As a result, zero copy techniques are implemented to reduce the unnecessary data copies and improve the CPU performance. Many different schemes are used to help implement zero copy techniques, such as memory remapping, shared buffers, different system calls, and hardware support.

3.1 Dynamic remapping

Dynamically remapping memory can eliminate the data copies between the application buffer and kernel buffer. As **Error! Reference source not found.** shows, firstly, data loaded from disk is stored in a kernel buffer by DMA copy. Then the pages of the application buffer are mapped to the kernel buffer, so that the data copy between kernel buffers and application buffers are omitted.

According to Stancevic [3], to remap the memory, the Linux `mmap()`system call will be called instead of the `read()` system call. After data are copied to the kernel buffer by DMA copy, the system call `write()` causes the data to be copied from the kernel buffer to another socket buffer in kernel address space. This means that the original copies from kernel buffer to application buffer and following copies from application buffer to socket buffer are substituted by only one data copy directly from kernel buffer to socket buffer. Then data will be copied to the network card buffer by DMA. Dynamic remapping needs three CPU copies (two DMA copies and one CPU copy) and four context switches, since the system calls `mmap()` and `write()` are made during data transmission.

In an environment requiring encryption, unless the kernel supports an encrypt-while-copy operation, we must allow the user application to copy data from the kernel buffer to the socket buffer. In a filtering situation we must allow the user application to process the data. Therefore, unless we can link the application buffer to both the kernel buffer and the socket buffer, and have the security routine read from one to the other, we will get no savings using dynamic remapping. We also have to make sure that no other process can get access to the kernel buffer and bypass the security routine.

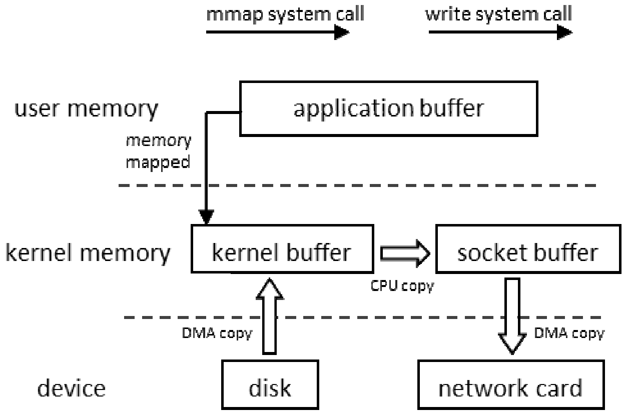


FIGURE 2: Data copy using memory remapping.

3.2 Shared Buffer in Kernel Memory Space

In INSTANCE-buffer [4], a memory area is shared in kernel space. As **Error! Reference source not found.** shows, there is a `buf` structure in the kernel buffer and `buf.b_data` is a pointer which points to the shared memory region which is in kernel memory space. Data fetched from disk are transferred to the shared memory area according to the `b_data` pointer. Furthermore, in the socket buffer, the m_data pointer in the `mbuf` structure points to the shared memory area. Therefore, data are copied from the memory pointed by `mbuf.m_data` to the buffer on the network card. According to Halvorsen and Jorde et al [4], the INSTANCE-buffer scheme allocates the shared memory region and the structures (buf and mbuf) statically, which means the time cost for allocating memory and variables is reduced.

In a secure environment, this technique will cause problems. To encrypt, we have to read the data from the source buffer and write it to a destination buffer. Even if we can encrypt in place, there will still need to be a read and write of data. A simple filter process will work as long as we can be sure that data is not sent out to the network card without the filter approving the message, but we cannot use a complex filter that modifies the data. In high-assurance systems this may not provide enough separation and isolation control.

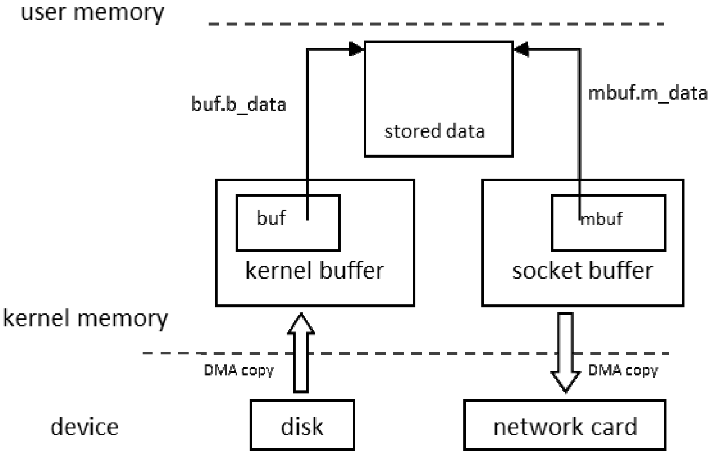


FIGURE 3: Shared buffer with structures.

3.3 Shared Buffer Between User and Kernel

In Linux, the sk_buff structure contains all the control information about individual network packets. Data sent to the network are copied to one or more sk_buff buffers. In LyraNET [5], the original sk_buff structure is modified to an enhanced copy elimination sk_buff structure, which eliminates the data copy by just passing the address of the data buffer. According to Chiang and Li [5], the modified sk_buff structure includes a new array with two elements, named dataseg, to record addresses for data without copying the data.

LyraNET is focused on reducing the data copies from user buffer to kernel buffer and from kernel buffer to network card. Using the enhanced copy elimination sk_buff structure, information about the protocol headers is copied to the network card and then data can be retrieved correctly according to the pointers and the data length stored in the modified sk_buff. LyraNET also uses DMA to copy data from the socket buffer to network card (Figure 4).

In a secure environment the security routine will have to copy data from the shared data region into a new region upon encryption. Filtering requires a CPU read of the data, but may not require a write. The system will have to ensure that the shared region is not available to the socket buffer until after it has been approved.

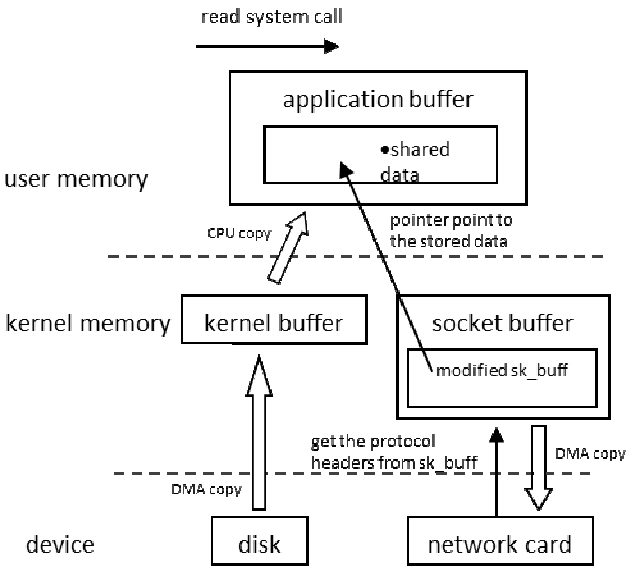


FIGURE 4: Shared user and kernel buffer

3.4 Sendfile

According to Stancevic [3], in Linux kernel version 2.1, the `sendfile()` system call was introduced to simplify the transmission of data over the network and between two local file descriptors. As Figure 5 shows, when calling the `sendfile()` system call, data are fetched from disk and copied into a kernel buffer by DMA copy. Then data are copied directly from the kernel buffer to the socket buffer. Once all data are copied into the socket buffer, the `sendfile()` system call will return to indicate the completion of data transfer from the kernel buffer to socket buffer. Then, data will be copied to the buffer on the network card and transferred to the network.

The `sendfile()` method is much more efficient than the combination of `read()` and `write()` system calls [6]. Compared with the traditional data transmission method, it replaces the `read()` and `write()` system calls, which reduces the number of context switch from four to two. Compared with dynamic remapping, `sendfile()` reduces the costs for virtual memory management. However, it still needs three data copies (two DMA copies and one CPU copy) to finish the data transmission.

This approach to zero copy bypasses the application completely, avoiding any application specific security routine or CPU processing of the data, making encryption and filtering impractical.

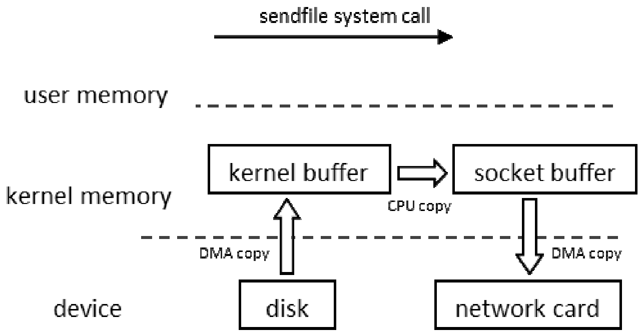


FIGURE 5: Data transfer using sendfile.

3.5 Sendfile With DMA Scatter/gather Copy

The method that enhances the `sendfile()` system call with DMA scatter/gather copy can eliminate the CPU copy between kernel buffer and socket buffer found in the `sendfile()` method. Different from the DMA which maps each buffer one by one and then do the operation, DMA scatter/gather copy maps the whole list of buffers at once and transfer them in one DMA operation [7]. In this method, support from hardware is needed, for the network interface gathers data from various memory spaces (Figure 6). When calling `sendfile(),` data on disk are loaded into a kernel buffer using DMA transfer. Then, only the buffer descriptor is sent to the socket, instead of coping all of the data to the buffer. The descriptor contains information about the length of the data and where it is. Using this information, header and trailer of the data packet can be generated. Then by using the DMA scatter/gather operation, the network interface card can gather all the data from different memory locations and store the assembled packet in the network card buffer.

Hardware which supports DMA scatter/gather copy eliminates the CPU copy between kernel buffer and socket buffer. It means that no CPU copies occur during the data transmission from disk to network card, which helps to increase the performance of the CPU. Only one system call, `sendfile(),` is made in this approach, so there are only two context switches. Additionally, since hardware supports the DMA scatter/gather operation, data in memory are not required to be stored in consecutive memory spaces.

This approach to zero copy bypasses the application completely, avoiding any application specific security routine or CPU processing of the data, making encryption and filtering impractical.

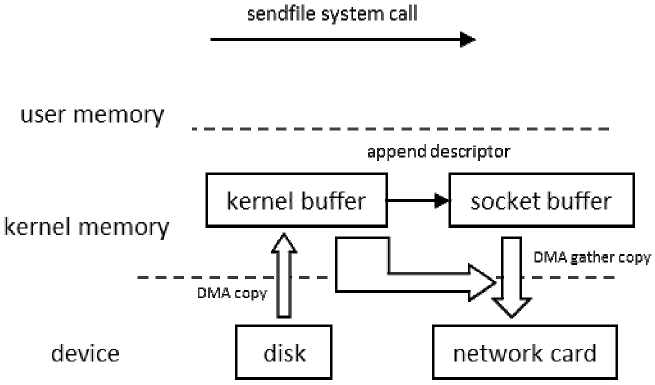


FIGURE 6: Sendfile and DMA scatter/gather copy.

3.6 Splice

According to the Linux man page [8], the Linux 2.6.17 kernel has a new system call, `splice()`. This call takes two file descriptors and an integer as parameters. It copies data, which size is specified by the integer, from one file descriptor to the other file descriptor using a pipe. By doing this, it does not need to copy data between kernel space and user space.

When using this approach, data are copied from disk to kernel buffer first. Then the `splice()` system call allows data to move between different buffers in kernel space without the copy to user space, for only file descriptors are transferred. In some zero copy implementations, `splice()`is called to move data from the kernel buffer to socket buffer which is also in kernel space. Then data are copied from the socket buffer to the network card buffer by DMA transfer. The system call splice eliminates the data copy between kernel buffer and socket buffer.

Unlike the method `sendfile()` with DMA scatter/gather copy, `splice()` does not need support from hardware. In addition, two context switches are bypassed compared to the traditional method, therefore only two context switches are needed. Instead of coping data to an application buffer, `splice()` works as a pipe between two kernel buffers to allow direct data transfer. However, there must be two file descriptors opened to both the input and output devices (Figure 7).

Fall and Pasquale [9] indicated that, in the best case, the performance of using `splice()` is at 1.8 times than the maximum throughput of the traditional data transfer method. This approach to zero copy bypasses the application completely, avoiding any application specific security routine, making encryption and filtering much more difficult.

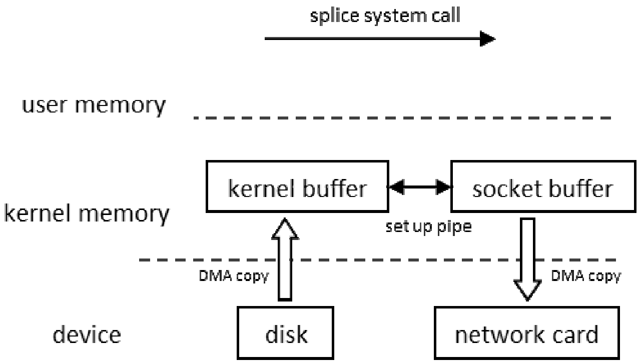


FIGURE 7: Data transmission using splice.

4. POTENTIAL PROBLEMS AND SECURITY CONCERNS

The implementation of zero copy technique helps to decrease overhead so as to release CPU cycles. Redundant data copies are reduced which saves more CPU and memory resources. However, the benefits of zero copy are sometimes overestimated, for the implementation of zero copy brings other new problems which may nullify the improvement of the system. For example, memory mapping needs costly virtual memory operations, the method of `sendfile()` with DMA scatter/gather copy needs support from hardware to gather data from various memory locations. In addition, zero copy techniques ignore or even introduce security concerns as well.

4.1 Dynamic Remapping Problem

Dynamic remapping uses `mmap()` instead of the `read()` system call, which reduce one copy during the data transmission. However, data might be modified by the user program while being prepared for transmission, this may cause security problems.

This problem can be prevented by implementing the Copy-on-write (COW) technique. COW is a simple technique. When the pages of the application buffers are mapped into kernel space, these pages can be written by the user. After the `mmap()` system call returns, both application and socket buffers can then only read the kernel buffer. The user can then start writing to the kernel buffer when the `write()` system calls return, which means the required data are sent to socket buffers. COW is implemented when using page remapping. It helps to ensure that while data is transferring, write permission is controlled. So the content of the physical memory page cannot be changed. However, COW is a costly event which consumes more CPU resource than a simple CPU copy [10].

4.2 Pre-allocated Buffer Problem

Some zero copy techniques use pre-allocated buffers to store data. By using pre-allocated buffer, the costly virtual memory operations are done only once. In addition, the pre-allocated buffer eliminates the exchange of the size of data between sender and receiver, but it causes another problem: the pre-allocated buffer space in receiver might be exhausted. If the receiver's pre-allocated buffers are exhausted, the receiver has to wait until some buffers are usable. According to Yamagiwa et al [11] even if there are a lot of pre-allocated buffers, the receiver still might be exhausted since these buffers cannot be paged out.

4.3 Sendfile System Call Problem

According to Brose [1], the `sendfile()` interface requires two file descriptors. The first one is a readable memory mappable file such as regular file or a block device. The second one can be a writable file or a network socket. The `sendfile()` system call can only transmit data which has specified length. The other disadvantage of the sendfile approach is that it is hard to implement `sendfile()` on the receivers' end, for network transfers are asynchronous and the `sendfile()` system call is not a standard system call among different Linux or UNIX systems. In some systems, `sendfile()` is routed to the `transferto()` system call, which makes the implementation of this approach more difficult.

4.4 Sendfile With DMA Scatter/gather Copy Problem

The method that uses `sendfile()` with DMA scatter/gather copy enhances the performance of the CPU by avoiding CPU copies. Since data can be copied from various memory areas to the network card buffer by DMA scatter/gather copy, which is supported by hardware, cache coherency is not required anymore. However, according to Brose [1], the source buffer may be a part of the page cache, which means it is available for the generic read functionality and may be accessed in traditional way as well. As long as the memory area can be accessed by the CPU, the cache consistency has to be maintained by flushing the caches before DMA transfers.

5. CONCLUSION

5.1 Zero-Copy Impacts

Over the past decade we have seen tremendous growth in e-government services at the state and federal level. The OMB report “FY 2009 Report to Congress on the Implementation of The E-Government Act of 2002” highlights federal agencies' efforts to implement e-government. These activities have included performance enhancing and cost saving consolidation of services into server farms and clouds, and increased availability of private and secure information. There are other reports that highlight similar efforts by the states. These reports indicate that there is a growth in the use and availability of e-government.

Vendors developing solutions for data servers will use new technologies if they can increase performance, such as zero copy. Although the implementations of zero copy techniques help reducing the redundant data copies between the kernel memory space and user memory space, they also bring us new problems. Memory remapping approach needs virtual memory operations and COW which might nullify the improvement of the CPU performance. When using the pre-allocated buffer or static buffer, size of the buffer becomes a problem. Because the buffer overflow attack and the condition that the buffers are exhausted need to be considered.

Therefore, the implementation of zero copy techniques regard to network is limited by many factors. That is why zero copy techniques have not been widely adopted in the operating systems. Although these new techniques need more evaluation, zero copy is worth analyzing and implementing carefully for critical e-government services. Removing redundant CPU copies can improve the performance of servers, and this is especially useful for commercial and government servers which require high efficiency. The advantages and disadvantages of different zero copy techniques are compared in

TABLE 1. Any zero-copy technique that bypasses the application will limit the ability to perform security processing. Therefore the `sendfile()` and `splice()` system calls are not recommended for use in applications that require security processing. The right most column indicates whether security processing can be used with each zero-copy technique, providing some performance improvement.

	CPU copy	**DMA copy**	**System call**	**Context switches**	**Advantages and disadvantages**	**App-level security processing**
Traditional data transfer	2	2	*read* *write*	4	Redundant data copies consume many CPU cycles. 2 system calls cause 4 context switches. Large consumption of memory bandwidth.	Yes
Memory remapping	1	2	*mmap* *write*	4	Needs costly virtual memory operations. Needs COW to be implemented which costs a lot.	Yes
Shared buffer (INSTANCE-buffer)	0	2		0	Buffers need to be released after using. Pre-allocated buffer reduces the exchange of the data size, since the buffer size is fixed. Pre-allocated buffer needs less virtual memory operations, for the virtual memory operations only used when creating the buffer.	Yes
LyraNET	1	2	*read*	2	Need costly virtual memory operations. LyraNET still needs data copy from kernel buffer to application buffer.	No
Sendfile	1	2	*sendfile*	2	Reduces the costs for virtual memory management. *Sendfile* hard to implement due to asynchronous network transfers. *Sendfile* system call is not a standard system call among different Linux or UNIX systems.	Yes
Sendfile with DMA scatter/gather copy	0	2	*sendfile*	2	Reduces the costs for virtual memory management. Needs hardware which supports DMA scatter/gather copy. *Sendfile* hard to implement due to asynchronous network transfers. *Sendfile* system call is not a standard system call among different Linux or UNIX systems.	No
Splice	0	2	*splice*	2	When using *splice* system call, there must has two file descriptors which opened to both the input and the output devices.	No

TABLE 1: Comparison of different zero copy techniques.

5.2 Understanding the Implications

To better understand the implications of security on performance improving solutions, we recommend that the acquisition officer evaluate the proposed solution using questions similar to the following list:

1) Will the server need to provide encrypted communication?
   a) Can the encryption be performed at the connection-level by dedicated hardware, or will you need application-level processing?

   b) Will the server need to provide multiple encryption services, supporting different algorithms and protocols?
2) Will the service need to provide application-level security processing/filtering on the data being serviced?
   a) Will that processing be service specific (same processing for all users) or user-specific?
   b) Will the service require multiple filtering processes?
3) Does the server need to support multiple concurrent services (such as cloud servers)?
4) Does the proposed solution support the level of encryption and application processing required?

The answers to these questions will help acquisition officers and developers better understand the expected use cases of the system and better evaluate the use of performance-improving technologies such as zero-copy. Any system that requires the examination or processing of the contents of the message, such as encryption, or filtering will require either a hardware solution, or application-level processing.

A hardware solution, such as a cryptographic hardware module, allows the use of all of the zero-copy techniques by passing the data packet directly to the cryptographic hardware. However, if the system supports multiple encryption services, a hardware solution may become impractical, thus forcing the system to rely on an application-level software solution.

Any software-based solution that requires access to the full data packet, cannot utilize the zero-copy features with a “no” in the final column of TABLE 1, Lyrannet, Sendfile with DMA scatter/gather copy and Splice. Other techniques will allow application level security processing.

We believe that it would be beneficial to develop a set of benchmark scenarios that address encryption and filtering technologies at the connection and application level and the recommend that the vendors provide performance data with respect to these benchmarks. This can be done to help address a wide-variety of proposed performance-improving technologies, which may not work well for security-enabled environments, contrary to vendor’s claims.

6. REFERENCES

[1] E. Brose. “ZeroCopy: Techniques, Benefits and Pitfalls,” http://kbs.cs.tu-berlin.de/teaching/ws2005/htos/papers/streams-zcpy.pdf, 2005, [last accessed May 30, 2012]

[2] C. Huang, C. Chen, S. Yu, S. Hsu, and C. Lin. “Accelerate in-line packet processing using fast queue,” in Proc. 2010 IEEE Region 10 Conference, 2010, pp. 1048–1052.

[3] D. Stancevic. “Zero Copy I: User-Mode Perspective,” *Linux Journal,* vol. 2003 no 105, Jan. 2003, pp. 3.

[4] P. Halvorsen, E. Jorde, K. Skevik, V. Goebel, T. Plagemann, “Performance Tradeoffs for Static Allocation of Zero-Copy Buffers,” in Proc. Euromicro Conference, 2002, pp. 138-143.

[5] M. Chiang and Y. Li, “LyraNET: A Zero-Copy TCP/IP Protocol Stack for Embedded Operating Systems,” in Proc. IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, Aug. 2005, pp. 123-128.

[6] Linux Programmer's Manual, http://www.unix.com/man-page/Linux/2/sendfile/ Feb. 2010.

[7] J.Corbet, A.Rubini, and G.Kroah-Hartman. “Memory Mapping and DMA” in *Linux Device Drivers*, third edition, O’REILLY, Feb, 2005, pp.450.

[8] Linux Programmer’s Manual, http://www.unix.com/man-page/Linux/2/splice/ Sep. 2009.

[9] K. Fall and J. Pasquale, “Exploiting In-Kernel Data Paths to Improve I/O Throughput and CPU Availability,” in Proc. Winter 1993 USENIX Conference, 1993, pp. 327-333.

[10] H.K.J. Chu. “Zero-copy TCP in Solaris,” Proc. USENIX Annual Technical Conference, 1996, pp. 253–264.

[11] S. Yamagiwa, K. Aoki, and K.Wada. “Active zero-copy: A performance study of non-deterministic messaging,” in Proc. International Symposium on Parallel and Distributed Computing, 2005, pp. 325–332.

Flow Modeling Based Wall Element Technique

Sabah Tamimi *sabah@agu.ac.ae*
Dean/College of Computing
Al Ghurair University
Academic City, Dubai, United Arab Emirates

Abstract

Two types of flow where examined, pressure and combination of pressure and Coquette flow of confined turbulent flow with a one equation model used to depict the turbulent viscosity of confined flow in a smooth straight channel when a finite element technique based on a zone close to a solid wall has been adopted for predicting the distribution of the pertinent variables in this zone and examined even with case when the near wall zone was extended away from the wall. The validation of imposed technique has been tested and well compared with other techniques.

Keywords: Pressure Flow, Combination of Pressure and Couette Flow, Expanding the Near Wall Zone.

1. INTRODUCTION

The Navier-Stockes equations governing fluids motion are known to apply in a wide range of applied computer science and engineering disciplines. Due to the complexity of these equations an analytical solution is intractable and during the last three decades, attention has been focused on the numerical simulation of flow process, the so called computational fluid dynamics (CFD). This has been developed and used with confidence to solve a large range of flow problems especially where experimentation is extremely difficult to obtain. It is known that when a fluid enters a prismoidal duct the values of the pertinent variables change from initial profile to a fully developed form, which is thereafter invariant in the downstream direction. The analysis of this region, which is known as developing region, has been the subject of extensive studies. Numerous theoretical and experimental works are available on laminar flow [1-4], but this is not the case of turbulent flow are still few since it has not been possible to obtain exact analytical solutions to such flows. Therefore, an effective technique is required to model the variation of the pertinent variables near a solid boundary, where the variation in velocity and kinetic energy, in particular, is extremely large near such surfaces since the transfer of shear form the boundary into the main domain and the nature of the flow changes rapidly. Consequently, if a conversational finite element is used to model the near wall zone (N.W.Z.), a significant grid refinement would be required. Indeed, in most situations this would be so fine as to be impractical.

Several solution techniques have been suggested in order to avoid such excessive refinement [5-7]. A more common approach is to terminate the actual domain subject to discretisation (main domain) at some small distance away from the wall, where the gradients of the independent variables are relatively small, and then use another technique to model the flow behavior in the NWZ. In this paper, a different near wall zone modeling techniques is used to simulate turbulent flow in a smooth straight channel.

2. GOVERNING EQUATIONS

The investigation of this paper is related to steady - state incompressible two dimensional turbulent flow of a Newtonian viscous fluid with no body forces acting. For such a situation, the Navier-Stokes (N-S) equations associated with this type are,

$$\rho\, u_j \frac{\partial u_i}{\partial x_j} = - \frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\mu_e\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)\right] \tag{1}$$

Where i,j= 1,2. u_i, p are the time - averaged velocities and pressure respectively, ρ is the fluid density, μ_e is the effective viscosity which is given by $\mu_e = \mu + \mu_t$, μ and μ_t are the molecular viscosity and turbulent viscosity, respectively. The flow field must satisfy the continuity equation, which may be written as:

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{2}$$

Equation (1) and (2) cannot be solved unless the turbulent contribution to μ_e be provided. The simplest model is via an algebraic formula [8] which has limited application and therefore this model is not adopted in the present work, but an alternative (Prandtl [9]-kolmogorov [10]) model is used in which,

$$\mu_t = C_\mu \rho\, k^{1/2} 1_\mu \tag{3}$$

Where k is the turbulence kinetic energy, 1_μ is the length scale which is taken as 0.4 times the normal distance from the nearest wall surface. The distribution of k can be evaluated by transport equation;

$$\rho u_j \frac{\partial k}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] + \mu_t \frac{\partial u_i}{\partial x_j}\left[\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right] - E \tag{4}$$

Where E= $C_D \rho k^{3/2} / 1_\mu$, μ_t / σ_k is the turbulent diffusion coefficient, σ_k is the turbulent prandtl or Schmidt number and C_D is a constant. The governing equations 1, 2 and 4 are called the one-equation (k-l) model. Within the main domain the governing equations have been discretised by using the standard finite element method [11] and Galerking weight residual approach is adopted to solve the discretised equations. The flow domain is divided into quadratic 8-noded elements used to define the variations in velocity and kinetic energy and kinetic energy, and 4-noded elements used for pressure. Within the near wall zone, either conventional finite element can be used, however an excessive mesh refinement was needed which is expensive in computer time and memory, or universal laws [12] to bridge from a solid boundary to the main domain (Figure 1). In the present work, a finite elements technique has been adopted, using one-dimensional normal to the wall (Figure 2).

3. BOUNDARY CONDITIONS

In the present work, two types of turbulent flow are considered. These are pressure and pressure plus Couette flow. In both, fully developed Dirichlet conditions are assumed on all variables upstream. No slip condition were imposed on solid boundaries and tractions updated downstream. Tractions are given by,

$$\tau_{x_1} = - p + \frac{\mu_e}{\rho}\left(\frac{\partial u_1}{\partial x_1}\right) \qquad x_1\text{- parallel to walls}$$

$$\tau_{x_2} = \frac{\mu_e}{\rho}\left(\frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2}\right) \qquad x_2\text{- normal to walls}$$

4. RESULTS AND DISCUSSION

Two examples were used to validate the imposed wall element technique was tested and comparisons made with other accepted techniques and experimental results [14] when fully developed turbulent flow is considered in a parallel-sided duct of width D, which is taken as 1.0, and L is the channel length. Compatible fully developed velocity and kinetic energy profiles were imposed as initial upstream values and outlet values from the previous iteration used as new approximation to the values at the inlet until a converged condition is satisfied. Different Reynolds number based upon the width of the channel of 12.000, 50.000 and 70.000 were considered.

The first example was concerned with an analysis of pressure flow where both walls of the channel are fixed. Figure 3 shows convergent velocity profiles at the outlet which clearly shows that the velocity values obtained by universal profiles have some discrepancy from those obtained from the advocated technique. Figures 4 shows the results obtained from the adoption of the presently advocated technique exhibits excellent agreement with the correct solution which resulted from the complete mapping. These are, superior to those obtained using universal laws. Figure 5 shows excellent agreement between the imposed technique and experimental results [14]. Figure 6 refer to the kinetic energy, which prove once more, the “correct” values are remarkably close to those obtained from the proposed technique.

The next stage was concerned with the validation of the wall element technique in an extended near wall zone when the interface located at 0.48D and 0.47D from the symmetric line as shown in Figures 7-8. These figures show the downstream velocity and kinetic energy. Obviously the results obtained from the adoption of 1-D elements in one direction is still the most advantageous owing to the number of elements used in the near wall zone.

The second example was concerned with an analysis of combining pressure and Couette flow, with the lower surface stationary and upper surface moving at a constant speed. Fully developed turbulent velocity profiles and turbulent kinetic energy distribution were obtained and presented in Figure 9 and 10, respectively, these show comparisons with universal laws and experimental results [14]. As conclusion, the validity of the wall element technique has been tested and approved previously [15-16]. In the present work, this validity has been tested again and approved again that the location of the near wall zone limit does not seem to affect the values of the pertinent variables. This is a distinct advantage over the universal law approach where strict limits must be placed on the location the interface, and once more, the results obtained from the adoption of the wall element technique are significantly better than those obtained using the universal laws, and compare favorably with experimental results.

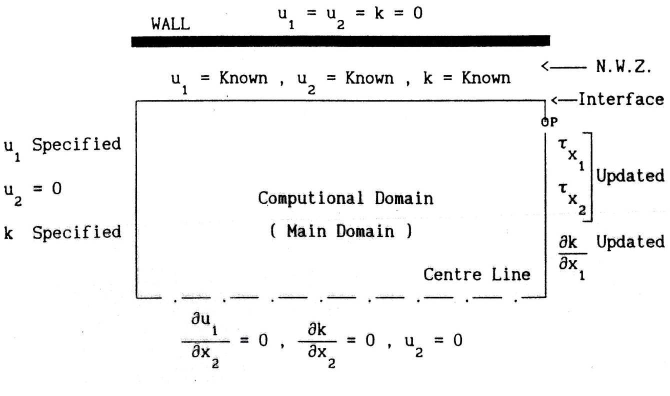

WALL
$u_1 = u_2 = k = 0$
u_1 = Known , u_2 = Known , k = Known
<—— N.W.Z.
<—Interface
P
u_1 Specified
$u_2 = 0$
k Specified
Computional Domain
(Main Domain)
Centre Line
τ_{x_1}
τ_{x_2}
Updated
$\frac{\partial k}{\partial x_1}$ Updated
$\frac{\partial u_1}{\partial x_2} = 0$, $\frac{\partial k}{\partial x_2} = 0$, $u_2 = 0$

FIGURE 1: Boundary conditions when the mesh is terminated at small distance away from the wall.

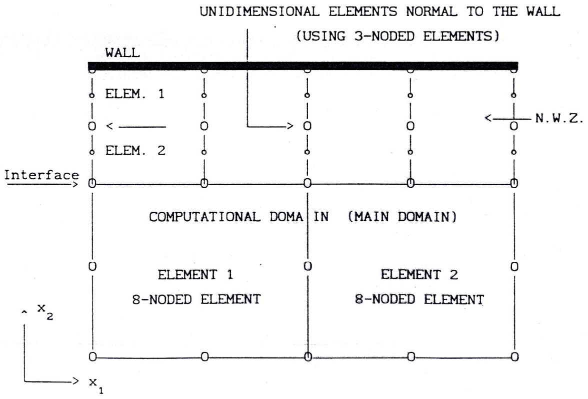


FIGURE 2: One-dimensional elements in one-direction normal to the wall used in the N.W.Z.

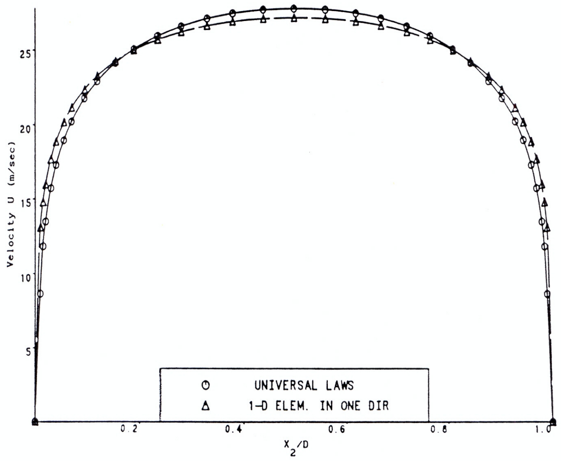


FIGURE 3: Turbulent velocity profiles for fully-developed flow, at flow, 8D downstream, L=8D, Re=50.000.

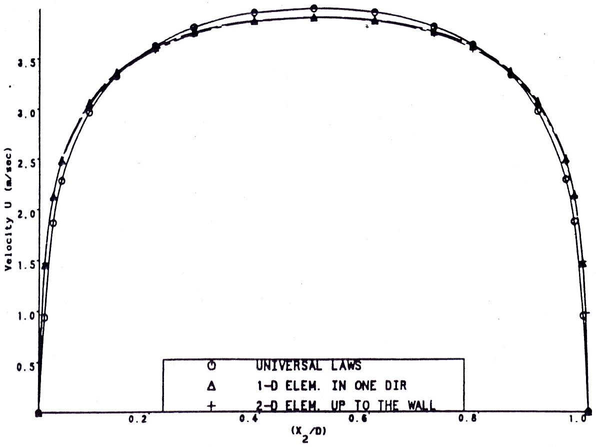


FIGURE 4: Turbulent velocity profiles for fully-developed at 8D downstream, L=8D, Re=12.000.

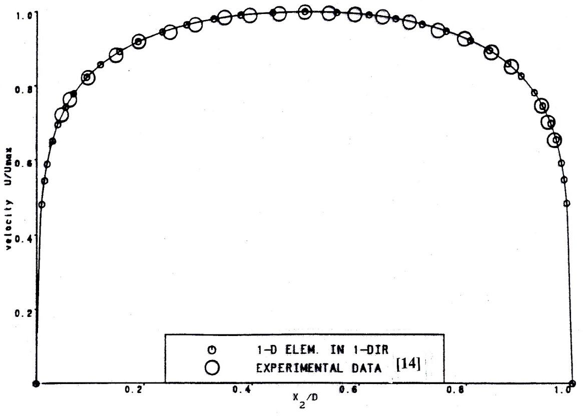


FIGURE 5: Turbulent velocity profiles for fully-developed flow, at 8D downstream, L=8D, Re=50.000.

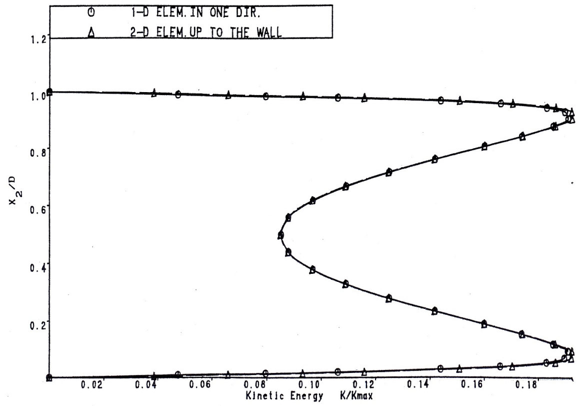


FIGURE 6: Kinetic energy profiles for fully-developed turbulent flow, at 8D downstream, L=8D, Re=12.000.

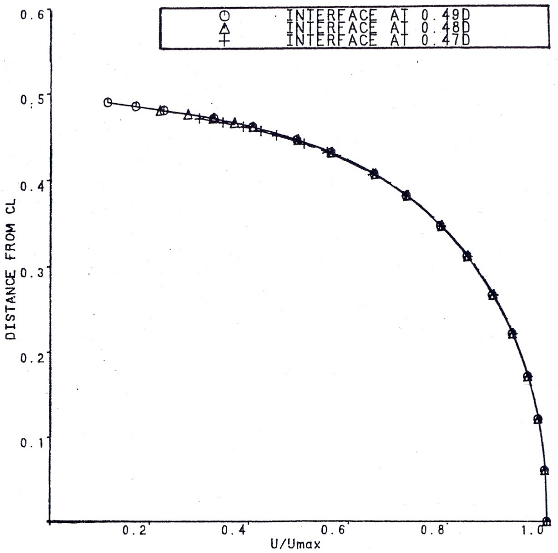


FIGURE 7: Downstream fully developed velocity profiles for turbulent flow when the N.W.Z. is extended up to 0.47D.

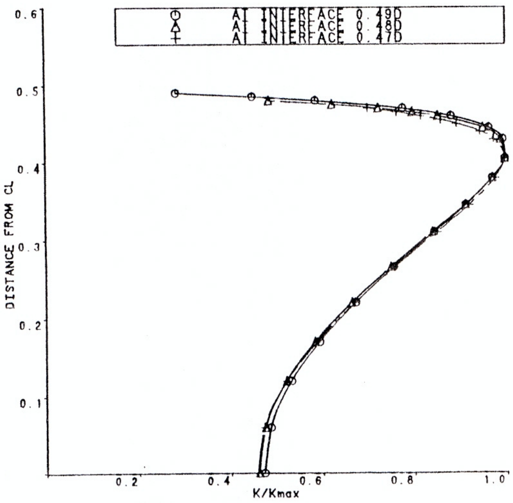


FIGURE 8: Downstream fully-developed kinetic energy profiles for turbulent flow when the N.W.Z. is extended up to 0.47D.

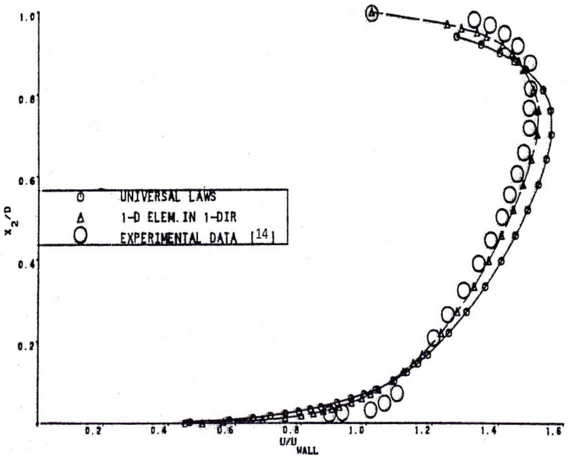


FIGURE 9: Velocity profiles for fully-developed for turbulent flow with fixed lower surface and moving upper surface, Re=70.000.

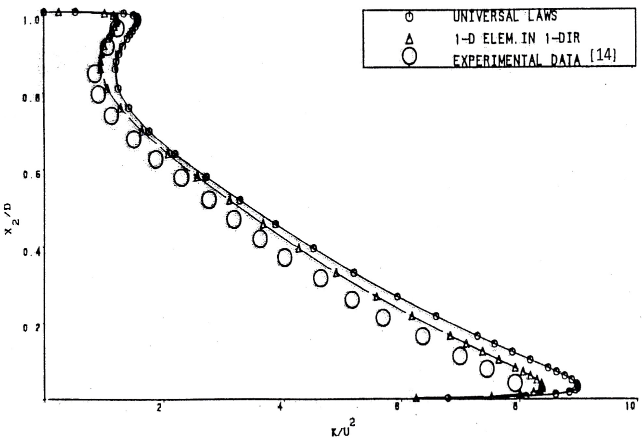


FIGURE 10: Fully-developed kinetic energy profiles for turbulent flow with fixed lower surface and moving upper surface Re=70.000.

5. CONCLUSIONS

The utilization of empirical universal laws is not valid since these laws are only really applicable for certain unidimensional flow regimes, and the general use of 2-D elements up to the wall is not economically viable. Therefore to avoid such an excessive refinement, these methods have been replaced by introducing a wall element technique, based on the use of the finite element methods which has shown an excellent results, when the fully-developed flow considered for both types of flow pressure and combination of pressure and Couette.

Again, the validation of the wall element technique in an extended near wall zone has shown more advantages comparing to the use of universal laws. Therefore, the imposed technique can be used with confidence for fully-developed turbulent flow.

6. REFERENCES

[1] C.L. Wiginton and C. Dalton, "Incompressible laminar flow in the entrance region of a rectangular duct", *J. Apple. Mech.*, vol. 37, 1970, pp. 854-856.

[2] E.M. Sparrow, C.W. Hixoin, and G. Shavit, "Experiments on laminar flow development in ractangular duct", *J. Basic Eng.*, vol. 89, 1967, pp. 116-124.

[3] D.M. Hawken, H.R. Tamaddon-Jahromi, P. Townsend and M. F. Webster, "A Taylor-Galerkin based algorithm for viscous incompressible flow", *Int. Journal Num. Meth. Fluids*, 1990.

[4] A.K. Mehrotra, and G.S. Patience, "Unified Entry Length for Newtonian and power law fluids in Laminar pipe flow", *J. Chem. Eng.*, vol. 68, 1990, pp.529-533.

[5] B.E. Launder, and N. Shima, "Second moment closure for near wall sublayer: Development and Application", *AIAA Journal*, vol. 27, 1989, pp. 1319-1325.

[6] Haroutunian, and S. Engelman, "On modeling wall-bound turbulent flows using specialized near-wall finite elements and the standard k-ε turbulent model", *Advances in Num. simulation of Turbulent flows, ASME*, vol. 117, 1991, pp. 97-105.

[7] T. Graft, A. Gerasimov, H. Lacovides, B. Launder, "Progress in the generalization of wall-function treatments", *Int. Journal for heat and fluid flow*, 2002, pp. 148-160.

[8] B.E. Launder, and D.B. Spalding, "Lectures in mathematical models of turbulence", *Academic Press*, 1972.

[9] L. Prandtl, "Uber ein neues forelsystem fur die ausgebildete turbulenze", *Nachr. Akad. Der wissenschafft, Gottingesn*, 1945.

[10] A.H. Kolmogrov, "Equations of turblent motion of an incompressible fluid", *IZV Akad Nauk, SSSR Ser. Phys*, vol. 1-2, 1942, pp. 56-58.

[11] C. Taylor, and T.G. Hughes, "Finite element programming of the Navier-Stokes equation", *Pineridge press*, 1981.

[12] T.J. Davies, "Turbulent phenomena", *Academic Press*, 1972.

[13] G.E. Schneider, G.D. Raithby, and M. Kovanovich, "Finite element analysis of incompressible fluid flows incorporating equal order pressure and velocity interpolation", *Proc. Int. Conf. Num. Meth. in laminar and turbulent flow*, Pentech Press, London, 1978, pp. 89-102.

[14] U.S.L. Nayak, and S.J. Stevens, "An experimental study of the flow in the annular gap between a long vehicle and a low close-fitting tunnel", *Report: Dept. of Technology, Loughborough University of Technology*, 1973.

[15] Sabah Tamimi," Representation of Variables of Confined Turbulent Flow in a Region Close to the Wall", *12th WSEAS International Conference on Applied Computer Science (ACS 12*), ISSN: 1790-5109, Singapore, 2012pp. 70-74.

[16] Sabah Tamimi, "Validation of Wall Element Technique of Turbulent Flow ", GSTF *Int. Journal on Computing,* ISSN: 2010-2283, Volume 2, No. 1, 2012, pp.177-181.

Using Geographical Location as an Authentication Factor to Enhance mCommerce Applications on Smartphones

Torben Kuseler *torben.kuseler@buckingham.ac.uk*
Applied Computing Department
The University of Buckingham
Hunter Street, Buckingham, MK18 1EG, UK

Ihsan Alshahib Lami *ihsan.lami@buckingham.ac.uk*
Applied Computing Department
The University of Buckingham
Hunter Street, Buckingham, MK18 1EG, UK

Abstract

Smartphones are increasingly used to perform mCommerce applications whilst on the move. 50% of all Smartphone owners in the U.S. used their Smartphone for banking transactions in the first quarter of 2011. This is an increase of nearly 100% compared to the year before. Current techniques used to remotely authenticate the client to the service provider in an mCommerce application are based on “static” authentication factors like passwords or tokens. The fact that the client is on the move, whilst using these mCommerce applications is not considered or used to enhance the authentication security. This paper is concerned with including client’s geographical location as an important authentication factor to enhance security of mCommerce applications, especially those requiring robust client authentication. Techniques to integrate location as an authentication factor as well as techniques to generation location-based cryptographic keys are reviewed and discussed. This paper further outlines restrictions of location as an authentication factor and gives recommendations about correct usage of client’s location information for mCommerce application’s authentication on Smartphones.

Keywords: Authentication, Location, mCommerce Applications, Security, Smartphone.

1. INTRODUCTION

Smartphones are becoming a major part in everybody’s daily life. All kinds of activities, including banking or financial mCommerce transactions (e.g. online shopping), are nowadays performed online via Smartphone applications whilst on the move. 50% of all Smartphone owners in the U.S. used their Smartphone for banking transactions in the first quarter of 2011. This is an increase of nearly 100% compared to the year before [1]. However, most of the techniques used to authenticate the client towards the remote authenticator (i.e. the bank offering a financial service) in these mCommerce applications still base upon classic (and static) authentication factors like passwords, tokens or biometrics. The fact that the client is on the move, whilst using these mCommerce applications is not considered or used to enhance the authentication security.

Reliable client authentication and data protection are still major concerns for mCommerce application providers because the classical authentication factors are open for hackers. As a result, mCommerce application providers restrict access, on average, to 30% of possible services to their clients via Smartphone applications [2].

This paper a) reviews techniques that use location as an authentication factor, and b) makes recommendations how location can be used to enhance the security of mCommerce applications requiring robust client authentication. This shall encourage mCommerce application providers to offer more services via Smartphone application to their clients.

The rest of this paper is organised as follows. Section 2 gives technical background information about methods to determine the location of Smartphones. Section 3 reviews techniques that use location as an authentication factor. In section 4, the use of location to generate cryptographic keys is discussed. Section 5 outlines restrictions of location as an authentication factor and makes recommendation towards a secure and reliable usage of location information in authentication. Finally, section 6 concludes the findings studied in this paper.

2. TECHNICAL BACKGROUND OF METHODS TO LOCATE SMARTPHONES

Three localisation techniques are commonly used to establish the location of Smartphones [3]. These techniques vary in the provided location accuracy (i.e. how exact can the technique determine the Smartphone’s location?) as well as the availability (i.e. does the technique cover the complete earth or only urban areas? Is the technique available indoors or does the client have to be outdoors to determine his/her position?).

2.1 Global Positioning System (GPS)

GPS-based positioning [4] has become the positioning technique mostly used on Smartphones. All new developed Smartphones feature a GPS receiver. GPS positioning is based on the reception of signals continuously transmitted from satellites. These signals contain the precise time the message was sent, as well as the location in orbit of the satellite. The GPS receiver uses the received signals of four or more satellites to calculate the current position based on trilateration. When outdoors, current GPS receivers onboard Smartphones are able to reduce the positional error to few meters [5]. However, GPS requires a line of sight to the satellites. Because of that, GPS can not be used (or the use is limited and the position becomes imprecise) indoors or in urban areas with many high glass-front buildings, where a direct line of sight to the satellites is not available. New satellite systems are rolling out such as GLONASS, and Galileo. These systems offer more enhanced signals and will provide better localisation accuracy than GPS.

2.2 Wi-Fi-based Positioning

Wi-Fi-based positioning uses Wi-Fi access points (Wi-Fi APs) to determine the position of the Smartphone. Wi-Fi APs continuously transmit beacons, including an AP identifier, to their surrounding area to inform potential Wi-Fi clients, such as a Smartphone, about their existence. Over the last years, several databases of APs and their corresponding geographical locations were collected by companies like Skyhook [6]. The Smartphone can use the AP identifier enclosed in the beacons and these databases, via an internet link, to determine the locations of the surrounding APs, by searching the identifier in the database. Depending on the number of APs in range, the achieved location accuracy of Wi-Fi-based positioning can vary between a few to 100 meters. Wi-Fi-based positioning can be used indoors as well as outdoors, as long as the AP transmitted beacon can reach the Smartphone. However, the number of available APs differs greatly between urban and rural areas, making Wi-Fi-based positioning a technique to be mainly used in big cities with lots of existing and known APs [5]. APs are also used to transport needed aiding information to the GPS device onboard the Smartphone. This helps the GPS receiver to fix much quicker.

2.3 Cellular Network Based Positioning

Cellular network based positioning use trilateration techniques to calculate the current Smartphone location [7]. The cellular network is divided into cells, in which each cell has a unique identifier (cell-ID). Depending on the trilateration technique used to determine the current phone location (e.g. U-TDOA [8]) and the cell size, cellular network based positioning accuracy can range between 50 metres to a few kilometres [5].

3. LOCATION AS AN AUTHENTICATION FACTOR IN REMOTE AUTHENTICATION

3.1 Classic Authentication Factors

Classic authentication factors are mostly used to authenticate the client towards the remote authenticator in mCommerce applications. Classic authentication factors can be categorised into three groups [9]:

1) Knowledge-based, or “something you know”
   Knowledge-based authentication factors rely on a memorised piece of information, e.g. PIN or password. Long and random passwords can offer a high level of security in authentication systems. However, in practice, clients have huge difficulties to memorise random and strong passwords. This often results in the use of short passwords that are therefore simple to guess and do not provide high authentication security.

2) Object-based, or “something you have”
   Object-based authentication factors rely on physical possessions, e.g. tokens. A token has the advantage over a knowledge-based authentication factor that clients do not need to memorise anything. This eliminates the risk of attackers guessing passwords easily because simple passwords are used. However, the main security drawback of physical tokens is that, when lost or stolen, an attacker gains unauthorised access.

3) Identity-based, or “something you are”
   Identity-based authentication factors, i.e. Biometrics rely on the uniqueness of physiological (e.g. fingerprint, facial features) or behavioural (e.g. hand-writing, speech) characteristics of a client. Biometric-based authentication offers two advantages over the other classic authentication factors:
   1) A client does not need to remember or carry anything.
   2) Biometrics verify the de facto client and not only knowledge of a password or possession of a token, i.e. the genuine client needs to be present at the biometric sensor.

   Biometric authentication systems are not perfect and their security can also be undermined. For example, the genuine client’s biometric can be replaced in the biometric template database or a biometric sample can by replayed by an attacker.

These classic authentication factors can be used to define the “**who**” of an authentication attempt. They define neither the “**where**” nor the “**when**” of the attempt, two similarly important properties of secure remote client authentication, for example to tackle distance or replay attacks. Thus, location (to define the “**where**”) and time (to define the “**when**”) should be integrated as further authentication factors, to define all these properties in remote client authentication.

3.2 Location as an Authentication Factor

Location was integrated into authentication systems as a factor to “ground” authentication attempts [9]. This “grounding” reduces the risk of distance attacks, because an attacker cannot claim to be at a location, the attacker actually does not is [10]. To achieve “grounding”, a unique identifier (digital signature) was derived from a GPS-based location and real-time on a specialised location signature sensor (LSS). The generated digital signature was then combined with further authentication data in such a way that it stamps the data with location and time information in a forgery-proof way. The security and uniqueness of the LSS digital signature bases upon the fact that bit values of GPS signals change every 20 milliseconds and so the resultant signature changes accordingly. However, current GPS receivers available on Smartphones cannot be used to generate such unique and trusted signatures, because these GPS receivers compute longitude and latitude from the received signals straightaway. Also, dedicated LSS are required to verify the client’s location signature. This aspect prevents that the LSS-based system can be deployed on a large scale, e.g. country wide, because this requires installation of thousands of LSS for verification of the client's claimed location. Thus, the LSS-based system is more suitable for limited areas like company premises but not for general mCommerce applications.

A similar approach with global availability is Secure Authentication for GPS phone Applications (SAGA) [11]. In contrast to other GPS-based services, SAGA can be used to determine the current location using the Smartphone's onboard GPS receiver as well as used to verify this claimed location. A security analysis of SAGA concluded that SAGA offers reliable and secure location verification, with the advantage that a GPS-based system is available worldwide [12]. However, to perform a verification of the client's claimed location, the SAGA system also requires additional trusted signal receivers at several known locations that are used to receive reference signals from the satellites for comparison. This introduces further costs for installation and maintenance of these receivers for practical authentication in mCommerce applications.

"Location cross-checking" techniques do not require additional receivers to be installed for location verification. Instead, location cross-checking compares the actual location of the Smartphone with a pre-agreed set of known points of businesses related to the registered clients (e.g. an ATM machine) to counter distance attacks [13]. However, location cross-checking requires an ongoing monitoring to track the current location of the Smartphone (to help identify abnormal activities or attacks to the system). This ongoing monitoring is difficult to maintain as Smartphones might be switched off by the client to save energy or might be used outside the traceable area. A further downside of location cross-checking lies in the dependence of the pre-agreed points of businesses, which are difficult to define and maintain in mCommerce applications for Smartphones.

"Location proofs" try to overcome the drawbacks of location cross-checking [14]. A location proof is a piece of data generated by a stationary sender (e.g. Wi-Fi AP) that is then sent to the Smartphone on request. The Smartphone stores the received proof for immediate or later use and attaches it to an authentication message to proof the client's current location. An advantage of location proofs over location cross-checking is that the points of businesses must not be defined in advance. However, location proofs require trusted stationary senders instead (e.g. Wi-Fi AP), which should not be easily susceptible to manipulation.

Location proofs are also critical from a client's privacy point a view [15]. Requesting a location proof discloses the client's identity to the stationary sender, i.e. the proof issuer. This information could then, for example, be used to generate a location profile of the client. To overcome this problem, the VeriPlace architecture [15] extended the location proof concept and included two separated and trusted entities for managing location and identity information of the client. This ensures that location and identity are never available at the same time to one entity.

The Privacy-Preserving Location proof Updating System (APPLAUS) [16] removed the requirement of stationary senders to issue location proofs. Instead other Smartphones in the close neighbourhood serve as location proof issuers and communicate the proof to the requestor via Bluetooth in a peer-to-peer approach. The benefit of APPLAUS is that no specific network infrastructure or specialised trusted senders are required. However, security and reliability of APPLAUS bases completely upon the number and trustworthy of the neighbouring Smartphones that issue the location proofs.Systems like APPLAUS may be adequate for low-value services that require a location proof, e.g. downloading a digital brochure of a museum for free if the client has previously visited the museum. For high-value mCommerce transaction authentication, such peer-to-peer architectures do not offer enough security and reliability, because the neighbouring Smartphones, which issue the location proofs, cannot be fully trusted.

Localisation and certification services [17] are used to tag digital content (e.g. authentication messages) with a location and timestamp DTL-certificate (Data-Location-Time). The DTL-certificate enables the receiver of the stamped content to verify where the content was originally created. To get a DTL-certificate, the client sends the hash value of the message to a localisation

/ certificate authority. This authority then determines the client's location, for example via cellular network based positioning or Wi-Fi-based positioning (cp. section 2). The determined location and current time are then combined with the hash value of the message and send back to the client as the DTL-certificate. To ensure the producer's privacy, the DTL-certificate does not include any information about the producer's identity. Thus, only time and location of the generated content can be verified by the receiver of the DTL-certificate.

The independent determination of the client's location by the localisation / certificate authority ensures that the client actually is at the claimed location, i.e. the DTL-certificate does not base upon the location determined by the client. However, the DTL-certificates miss a tight binding between location and client. The generated DTL-certificate can be given away and used by others, which could undermine the authentication system, i.e. an attacker could use a stolen certificate to impersonate a genuine client's location. In addition, DTL-certificates might be subject to malicious modifications, because the DTL-certificates have to be sent back to the client and are stored on the client's phone.

4. LOCATION-BASED KEYS

Section 3 reviewed and discussed approaches that use location information as an authentication factor directly integrated into an authentication message to "ground" the client's authentication attempt. Another possibility to use location information to enhance the security of mCommerce applications is to combine location with established cryptographic algorithms, i.e. message encryption. For example, the authenticator can encrypt his authentication messages to the client with a cryptographic key based on a combination of a) a client specific and pre-agreed password and b) a location-based key. This has the advantage in mCommerce application that an attacker needs to get two pieces of information (i.e. the genuine client's password and the current location) to illegitimately decrypt the authenticator's messages.

The GeoEncryption concept [18] utilises this approach and uses location as a source to generate location-based keys for encryption of digital messages. GeoEncryption extends a classic hybrid cryptographic algorithm with a GeoLock functionality to ensure a secure location binding of the digital message. A geo-encrypted message can be opened successfully (decrypted) by a receiver, if the receiver's actual location is inside the required area. A general GeoLock mapping function, based on the estimated Position, Velocity, and Time (PVT) on the recipient is used to generate the message encryption and decryption key. However, the GeoEncryption concept did not specify a practical and secure PVT mapping function nor does it handle support of mobile and moving recipients, which is an important aspect of mCommerce applications performed on Smartphones.

In the added GeoEncryption mobility model [19], the encrypted message receiver continuously updates the sender about his/her current location. The sender then uses this information to dynamically adjust the decryption area in which the receiver can decrypt the message. A practical mapping function for the GeoEncryption concept uses square areas [20]. This mapping function was then improved to cover any shape [21], which increased the precision of how the decryption area can be specified, i.e. the introduced decryption area error is reduced.

A drawback of these mapping functions is that the generated encryption key merely bases upon the geographical coordinates (longitude and latitude values) of the decryption area and the used hash function as shown in Figure 1 [20]. If the geographical coordinates of the target region (decryption area) can be estimated by an attacker, because the attacker is in close proximity to the recipient, then the complete security of GeoEncryption lies in the secrecy of the hash function. If the used hash function is also known to the attacker, then the attacker is able to decrypt the message. To minimise this risk, the GeoEncryption keys should be combined with further client specific information (e.g. password or a token stored on the client's Smartphone) that is more secret than the "public available" location of the client (cp. section 5).

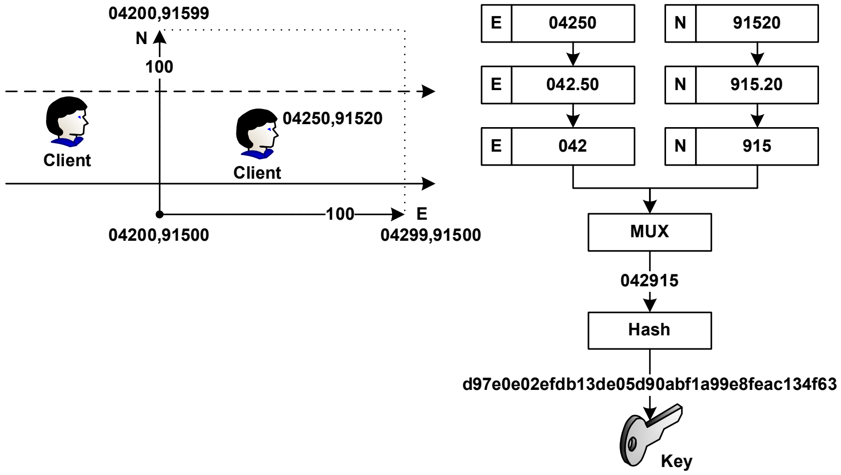


FIGURE 1: GeoEncryption mapping function

Time and Location Based One-Time-Passwords (TLB OTP) [22] utilises the estimated location of the client with the current time to calculate a TLB OTP. This TLB OTP is then used as a key to (de)encrypt all further communication messages between client and authenticator. Adding location to classical OTP schemes, which are merely time-dependent, strengthens the authentication security, because it is more difficult for an attacker to determine the client's current location and precise time simultaneously [22]. To enhance and ensure correctness of the client's location and future position estimation, the client sends periodically update information about the client's current location and movement to the authenticator.

The practical mapping functions for GeoEncryption also require that the recipient's direction of movement is known during encryption of the message to correctly define the decryption area. To achieve this, the receiver also transmits periodically movement updates, which are then used to calculate the correct decryption area [20]. The importance of these updates can be seen in the example of Figure 1. The starting point of the decryption region is chosen to be at: "E04200" and "N91500", and the client is assumed to travel eastwards (E) in a maximum range of 100 meters (i.e. location will be equal to "E04299" after 100 meters). If the client travels more than 100 meters, a different key is produced. For example, travelling 110 meters results in "E04310" and hence the hashed value is completely different. A similar problem occurs if the direction of the client's movement is unknown. In this case, the phone needs to move one meter westwards instead of eastwards (i.e. to location E04199) to produce a different key. Such precise client location estimation is difficult to achieve in mCommerce applications, because clients often change directions when using their Smartphone's whilst on the move.

The Location-Dependent Data Encryption Algorithm (LDEA) introduces a Toleration Distance (TD) during encryption of the messages to overcome the receiver's movement uncertainty [23]. The TD shall guarantee that always the same key is generated on sender and receiver side, if the receiver is within the TD area. However, analyses of the LDEA showed that LDEA is not able to generate always the correct decryption key even if the client is within the specified TD area as shown in Figure 2.

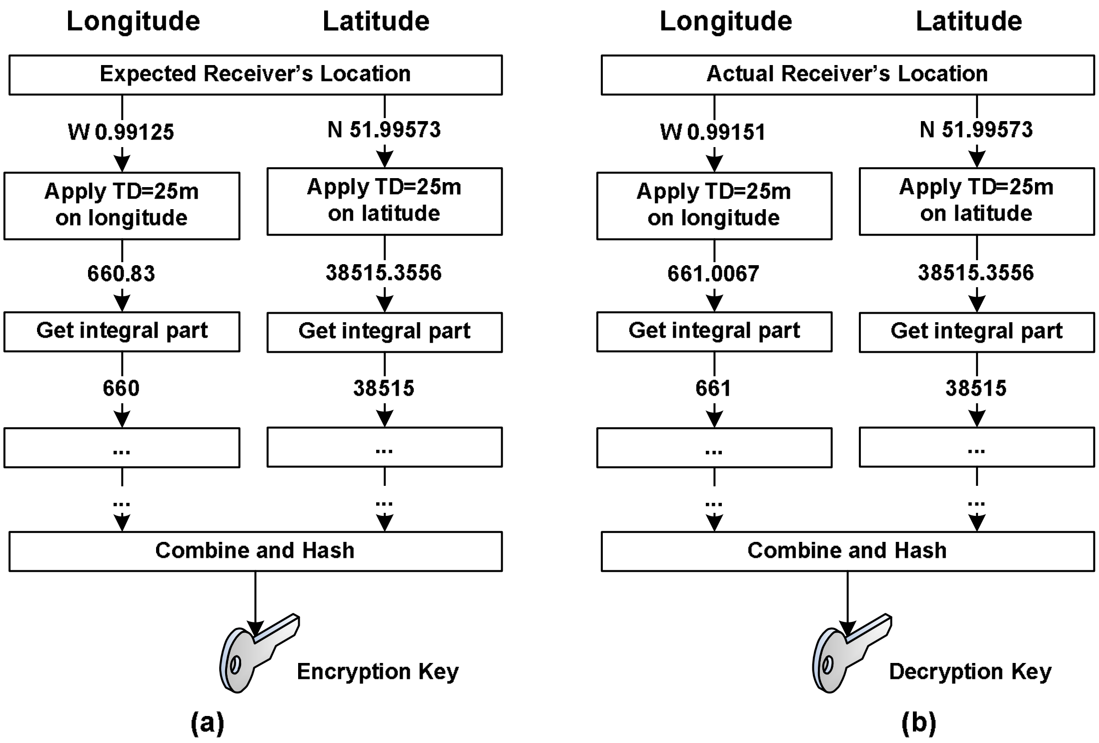


FIGURE 2: GeoEncryption key generation function

Figure 2(a) shows the process to generate the encryption key performed by the sender for the expected receiver's location of: W0.99125 / N51.99573 (cp. [23] for the details of this process). A TD of 25 metres is used in this example. Figure 2(b) shows the same process performed by the receiver on his/her actual location (W0.99151 / N51.99573). These two locations (i.e. expected and actual) are 17 metres away from each other and therefore well within the TD of 25 metres. However, the integral part of the longitude values is different (660 != 661). This means that the decryption key will also differ and that the client is not able to decrypt the message.

5. CONSIDERATIONS FOR USING LOCATION AS AN AUTHENTICATION FACTOR IN mCOMMERCE APPLICATIONS

5.1 Restrictions of Location as an Authentication Factor

Location as an authentication factor has restrictions compared to classical authentication factors (e.g. passwords, tokens or biometrics) and requirements, if location is used to generate cryptographic keys:

1) Location of a Smartphone is "publicly" available knowledge. Location can be easier gathered by an attacker, which is more difficult, for example, for undisclosed password. Attackers could simply follow clients and use the knowledge of the clients' whereabouts to get unauthorised system access, if location is the only factor used in the authentication system.

2) Use of location to generate cryptographic keys needs an appropriate key-generation function to transfer the physical client's location into a key. Utilising an inadequate transfer function can result in simple to guess location-based keys.

   It is important that the generated location-based key does not directly relate to the client's physical location (e.g. latitude and longitude values), i.e. a key for any location should not be predictable from a known location / key pair. If this property is not satisfied, then:

1) Large areas of the earth (e.g. Arctic, Antarctic) must be eliminated from the available key-space area, because it is unlikely that a client is in these areas.
2) The number of keys an attacker needs to try in a brute-force attack reduces tremendously, if the client's location is approximately known.

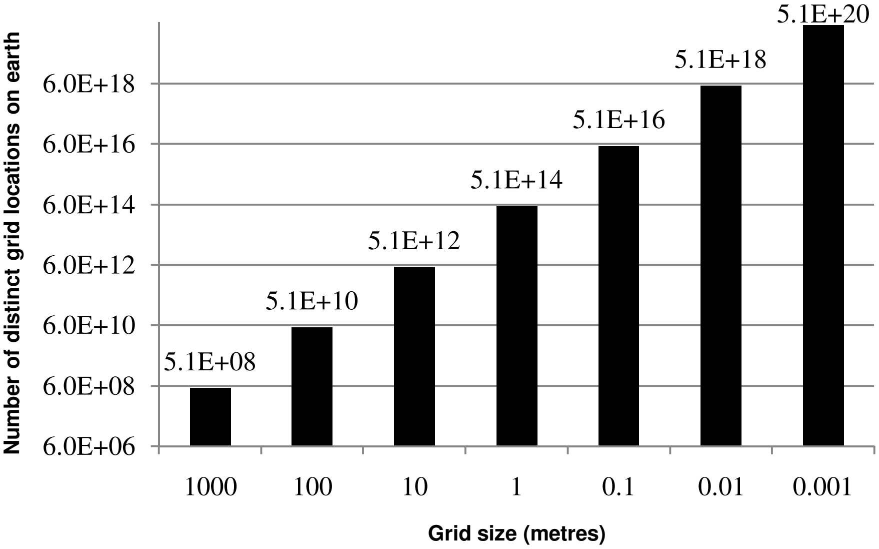


FIGURE 3: Maximum number of possible location-based keys

3) The number of locations on earth is limited. This restricts the number of possible locations to be used as location-based keys, i.e. the location key-space is restricted. Figure 3 shows the maximum number of location-based keys possible on earth, if the earth's surface is divided into a grid of equal-distance squares [24]. For a square size of one metre, $5.1*10^{14}$ different keys can be generated. This number is comparable to the number of eight character (0-9a-zA-Z) long passwords that is $2.2*10^{14}$. However, depending on the technique used to determine the clients location (cp. section 2), the location key-space can be much less, because the location determination technique does not achieve such a high accuracy.

5.2 Recommendations for Location as an Authentication Factor

To use location as an authentication factor in mCommerce applications or to generate secure and reliable location-based keys, the following recommendations should be applied:

1) The location-based keys should be determined completely independent from each other, without the need of any further sending of location information or movement direction to establish the same key on the client and authenticator side. This condition is required to guarantee that the authenticator can completely independently verify and consequently trust the claimed location of the client [25].

2) The location-based key needs to be generated with a specific location tolerance. This tolerance is necessary because methods to determine and verify the claimed client location differ in the accuracy (cp. section 2). A tolerance area should be used to handle this difference.

3) All location-based keys outside the tolerance area need to be different to the key representing the tolerance area. This condition ensures that the client is actually inside the tolerance area and not at a different place, which may produce the same key.

4) It must be ensured that the key-space of the generated location-based keys is large enough to minimise the risk of a brute-force attack (cp. section 4). This can be achieved, for example,

be combining client's location with further authentications factors (e.g. passwords) in the key generation process.

5) The location-based key should incorporate location information as well as further, more secret data (e.g. a token stored on the client's Smartphone). This eliminates the risk that an attacker is able to calculate the location-based key, if the attacker knows the client's current whereabouts.

6) Introduction of client's physical location into the authentication process may raise "privacy concern", i.e. tracking clients' location without their consent. To overcome this concern, methods which preserve the client's location privacy [26] and, at the same time, enable the authenticator to verify the client's claimed location independently should be used.

6. CONCLUSION

Integration of geographical location of the client's mobile device as an authentication factor into remote authentication systems for mCommerce application shall enhance the security of such systems:

1) Remote attacks are reduced, because integration of location information into the authentication data "grounds" the authentication attempt to a specific place. If the client's location claim is then independently verified, then an attacker cannot pretend to be at a different place. Independent location verification is important, because the location determined on the phone can be manipulated. For example, the GPS receiver of the phone can be manipulated or an IP-address-based location determination can be fooled by using a proxy server. I.e., for example, a cellular network operator based localisation is used to verify the client's claimed or the GPS predicted location.

2) If the client's location is combined with real-time, then remote replay attacks can also be reduced. Client's authentication data can be uniquely stamped with the current time. I.e. an attacker cannot re-use previously gathered genuine client authentication data, because of the time-stamp expiry.

However, the use of location also introduces requirements (e.g. client's location privacy or limited location-based key-space), which need to be carefully addressed by the authentication system. Privacy preserving algorithms can be deployed to solve such issues. For example, the client's actual location is randomly projected based on the Cell-ID serving the client's mobile device.
Research on secure and reliable integration of location information into authentication as well as generation of location-based key is still ongoing and needs further investigations and improvements to widely deploy location as an authentication factor in mCommerce application for Smartphones.

7. REFERENCES

[1] Frank Diekmann, "Survey: Mobile Bankers Double Over Last Year." Credit Union Journal, vol. 15, no. 18, pp. 19-19, May 2011.

[2] Security's Role in Deploying Transaction-Enabled Mobile Applications, Aug 2010.

[3] G. Sun, J. Chen, W. Guo, and K.J.R. Liu, "Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs." IEEE Signal Processing Magazine, vol. 22, no. 4, pp. 12-23, 2005.

[4] U.S. Government,"Official U.S. Government information about the Global Positioning System (GPS) and related topics."Internet: www.gps.gov, Apr. 20, 2012 [May 15, 2012].

[5] Paul A. Zandbergen, "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning." Transactions in GIS, vol. 13, no. s1, pp. 5-25, 2009.

[6] Skyhook, “Skyhook.”Internet: www.skyhookwireless.com, [May 15, 2012].

[7] Axel Kuepper. Location-Based Services: Fundamentals and Operation. Wiley Online Library, Oct. 2005.

[8] TruePosition, U-TDOA: Enabling New Location-Based Safety and Security Solutions, Oct. 2008.

[9] S.Z. Li and A.K. Jain.Encyclopedia of Biometrics. US, Springer US, 2009.

[10] D. Denning and P. MacDoran, "Location-Based Authentication: GroundingCyperspace for Better Security." Computer Fraud and Security Bulletin, Feb. 1996.

[11] A.I.G.T. Ferreres, B.R. Alvarez, and A.R. Garnacho, "Guaranteeing the authenticity of location information." IEEE Pervasive Computing, pp. 72-80, 2008.

[12] S. Lo, D.S. De Lorenzo, P.K. Enge, D. Akos, and P. Bradley, "Signal authentication-a secure civil gnss for today." inside GNSS, vol. 4, no. 5, pp. 30-39, 2009.

[13] G. Becker, S. Lo, D. De Lorenzo, P. Enge, and C. Paar, "Secure Location Verification." Data and Applications Security and Privacy XXIV, 2010, pp. 366-373.

[14] A. Haeberlen et al., "Practical robust localization over large-scale 802.11 wireless networks." in Proceedings of the 10th annual international conference on Mobile computing and networking, ACM, 2004, pp. 70-84.

[15] S. Saroiu and A. Wolman, "Enabling new mobile applications with location proofs." Proceedings of the 10th workshop on Mobile Computing Systems and Applications, New York, USA, 2009, pp. 3:1--3:6.

[16] W. Luo and U. Hengartner, "VeriPlace: a privacy-aware location proof architecture." Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2010, pp. 23-32.

[17] Z. Zhu and G. Cao, "Applaus: A privacy-preserving location proof updating system for location-based services." INFOCOM, 2011 Proceedings IEEE, 2011, pp. 1889-1897.

[18] V. Lenders, E. Koukoumidis, P. Zhang, and M. Martonosi, "Location-based trust for mobile user-generated content: applications, challenges and implementations." Proceedings of the 9th workshop on Mobile computing systems and applications, ACM, 2008, pp. 60-64.

[19] L. Scott and D.E. Denning, "A location based encryption technique and some of its applications." in ION National Technical Meeting, vol. 2003, 2003, pp. 730-740.

[20] A. Al-Fuqaha and O. Al-Ibrahim, "Geo-encryption protocol for mobile networks," Computer Communications, vol. 30, no. 11-12, pp. 2510-2517, 2007.

[21] G. Yan and S. Olariu, "An efficient geographic location-based security mechanism for vehicular adhoc networks." IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, MASS'09, 2009, pp. 804-809.

[22] G. Yan, J. Lin, D.B. Rawat, and W. Yang, "A Geographic Location-Based Security Mechanism for Intelligent Vehicular Networks." Intelligent Computing and Information Science, pp. 693-698, 2011.

[23] W.B. Hsieh and J.S. Leu, "Design of a time and location based One-Time Password authentication scheme." Wireless Communications and Mobile Computing Conference (IWCMC), 7th International, IEEE, 2011, pp. 201-206.

[24] H.C. Liao and Y.H. Chao, "A new data encryption algorithm based on the location of mobile users." Information Technology Journal, vol. 7, no. 1, pp. 63-69, 2008.

[25] L. Scott and D.E. Denning, "Location Based Encryption & Its Role In Digital Cinema Distribution." Tech. rep. 2003.

[26] Ihsan A. Lami, Torben Kuseler, Hisham Al-Assam, and Sabah Jassim, "LocBiometrics: Mobile phone based multifactor biometric authentication with time and location assurance," Proc. 18th Telecommunications Forum, IEEE Telfor, Nov. 2010.

[27] Torben Kuseler, Hisham Al-Assam, Sabah Jassim, and Ihsan A. Lami, "Privacy preserving, real-time and location secured biometrics for mCommerce authentication," SPIE Mobile Multimedia/Image Processing, Security, and Applications 2011, vol. 8063, Apr. 2011.

Comparative Analysis of Algorithms for Single Source Shortest Path Problem

Mrs. Shweta Srivastava *shwetasrivastava21@gmail.com*
Computer Science & Engineering Department,
ABES Engineering College Ghaziabad.
India

Abstract

The single source shortest path problem is one of the most studied problem in algorithmic graph theory. Single Source Shortest Path is the problem in which we have to find shortest paths from a source vertex v to all other vertices in the graph. A number of algorithms have been proposed for this problem. Most of the algorithms for this problem have evolved around the Dijkstra's algorithm. In this paper, we are going to do comparative analysis of some of the algorithms to solve this problem.

The algorithms discussed in this paper are- Thorup's algorithm, augmented shortest path, adjacent node algorithm, a heuristic genetic algorithm, an improved faster version of the Dijkstra's algorithm and a graph partitioning based algorithm.

Keywords: Single Source Shortest Path Problem, Dijkstra, Thorup, Heuristic Genetic Algorithm, Adjacent Node Algorithm.

1. INTRODUCTION

The single source shortest path problem can be defined as: given a weighted graph (that is, a set *V* of vertices, a set *E* of edges, and a real-valued weight function *f*: *E* → **R**), and one element s of *V* (i.e. a distinguished source vertex), we have to find a path *P* from s to a *v* of *V* so that

$$\sum_{p\in P} f(p)$$

is minimal among all paths connecting *s* to *v* .
Refer figure 1.

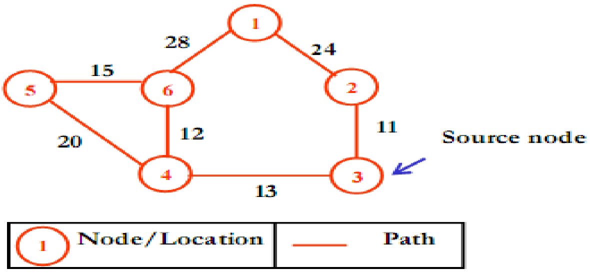


FIGURE 1: An Undirected Graph of 6 nodes and 7 edges

SSSP is applied in various areas such as:
1. Road Network
2. Computer Network
3. Web Mapping

4. Electronic Circuit Design
5. Geographical Information System (GIS)

In all six papers [1][2][3][4][5][6], the authors have given an improvement over the Dijkstra's algorithm.

2. BACKGROUND STUDY AND ANALYSIS

The Dijkstra's algorithm makes assumption that there is no negative-weight edges in the graph G (V, E) : w(u, v) >0 , ∀ (u, v) ∈ E. The algorithms in [1][2][3][4][5][6] also follows this assumption. For simplicity, we use n = |V| and m = |E| and K distinct edge lengths.

Several algorithms have been proposed for this problem which is based on different design paradigms, improved data structure, parameterization and input restrictions. According to survey it is found that no algorithm based on the Dijkstra's algorithm has achieved the linear time complexity due to drawbacks of Dijkstra's algorithm. Dijkstra's algorithm maintains an adjacency matrix which consumes n*n space in the memory. When the number of nodes (n) is very large, it is difficult to apply Dijkstra's algorithm.

So, in paper [1], an algorithm has been proposed which modify Dijkstra's algorithm to overcome its bottlenecks. Dijkstra's algorithm visits the vertices corresponding to a sorting algorithm (in order of increasing d (v)). Since there is no linear ime algorithm for sorting problem. Unless the order of visiting vertices is not modified, a linear time complexity cannot be achieved. The performance of an algorithm for single source shortest path problem depends on the 3 attributes:

(1) Preprocessing time: time required to construct a search structure suitable for search.
(2) Space: storage used for constructing and representing the search structure.
(3) Search Time: time required to find shortest path from a query source s, using the search structure.

In year 2000 **Thorup** proposed a concept of components and using some complicated data structures which overcome the problem in Dijkstra's algorithm. There are 3 interesting features of the Thorup's algorithm [1]:
(i)It contains a minimum spanning tree algorithm as its sub procedure. To achieve the linear time complexity Thorup used a linear time MST algorithm.
(ii) Thorup's algorithm consists of 2 phases: a construction phase which constructs a data structure suitable for a shortest path search from the given query source s; a search phase of finding the shortest paths from s to all vertices using the data structure constructed in construction phase.
(iii) Construction phase in Thorup's algorithm is independent of the source, while data structures in previous algorithms heavily depend on the source.
Summary of the Thorup's Algorithm:
Step1. Construct an MST (M).
Step2. Construct a component tree (T) using MST.
Step3: Compute widths of buckets (B) to maintain the components.
Step4: Construct an interval tree (U) to store unvisited children.
Step5: Visit all components in T by using B and U. Also known as search phase.
The running time of each step is as follows: step1: O (m) time, step2, 4: O (n) time, step5: O (m+n) time.

In year 2000, a linear time algorithm for SSSP problem [4] was proposed called an improvement over

Augmented Shortest Path Algorithm.
They proposed this algorithm for situations where no edge is unreasonably larger than the other edge and the ratio of maximum and minimum weights of the edges (f) of the graph is not very large.

The algorithm converts the graph G into an augmented graph (Ga) in we replace every edge in the original graph by number of edges having equal weights and number of new edges for each edge in the graph is bounded. Then the shortest path tree is obtained. The major drawback in the ASP algorithm was that if some heavy weight edge is included in the shortest path tree ASP fails to perform well so an improvement is done in improved ASP algorithm [4].

```
New_Augmented_Shortest_Path (G,s)
Find the minimum edge weight $w_{min}$ in O(m) time from the adjacency list.
for all (u,v) ε E do
 inqueue [u,v] <- false
 edge [u,v] <- ∞
end do
d[s]<-0
nowmin=∞
nextmin=∞
 nowcount = 0
for all (s,u) ε  E do
   enqueue (u,s)
   inqueue[s,u] <- true
   edge [s,u] <- w[s,u]
   if (w[s,u] < nowmin)
   nowmin=w[s,u]
  nowcount = nowcount+1
end do
nextcount = 0
while queue != empty do
  v,p <- serve()
  wv<- edge[p,v]-max(wmin , nowmin)
  if w <= 0 then
    if d[p]+w[p,v] <d[v] then
     d[v] <- d[p]+w[p,v]
     ∏[v] <- p
     for all (v,u) ε E do
        if inqueue[v,u] = false then
          if d[v]+w[v,u] < d[u] then
         enqueue(u,v)
         inqueue[v,u] <- true
         edge[v,u] <- w[v,u]+wv
         nextcount = nextcount + 1
         if(egde[v,u] < nextmin)
               nextmin = edge[v,u]
      endif
    endif
    else
      edge[v,u]<-min(edge[v,u],w[v,u]
      + wv )
      if(edge[v, u] < nextmin)
            nextmin = edge[v,u]
      endif
    endif
   enddo
  endif
  endif
  else
  if d[p]+w[p,v] <d[v] then
    enqueue(v,p)
    edge[p,v] <-wv
    if(wv   < nowmin)
        nextmin = wv
    endif
  endif
endif
nowcount=nowcount-1
if (nowcount==0)
 nowcount=nextcount
 nowmin=nextmin
 nextmin = ∞
 nextcount = 0
enddo
Return the shortest path.
```

This new algorithm proposes that the distance order search will advance by the maximum of w(min) and the minimum weight in the queue. So, the problem of larger weight edge got removed in this new version. The new augmented shortest path algorithm takes O(mf/2 + n) time which is linear if f is not large. The space requirement is O(m+n). This algorithm proved to perform better than the bucket based algorithm.

In year 2006, an approach was proposed by the authors of [6] to speed up the Dijkstra's algorithm. An acceleration method called arc-flag is used to improve Dijkstra's algorithm. In this approach we follow a preprocessing of the graph to generate some additional information about the graph which is then used to speed up shortest path queries. In the preprocessing step graph is divided in the regions and checked whether an arc belongs to the shortest path in the given region. This preprocessing method is combined with an appropriate partitioning technique and bi-directed search which achieves an average speed up factor of more than 500 compared to the Dijkstra's algorithm on large networks. They tested different combinations of the arc-flag method with different partitioning technique.

They used A*, bi-directed search techniques and chosen bi- directed search because A* didn't improve the speed up factor. They considered Grid, kd, Tree or METIS as the base partitioning method and made 11 combinations of the searching, partitioning and preprocessing techniques. They applied the 11 different combinations on the German road network data. Kd trees and METIS yields the best speed- up. Bi-directed search proved to be better than the unidirected search and the two level partitioning was better than the single level partitioning. The preprocessing takes O (m(m+n+nlogn)) time. It increases for the dense graph.

In year 2007, **a heuristic genetic algorithm** [3] was proposed to achieve high performance. Their proposal starts with the initial population of candidate solution paths than a randomly generated one. HGA also uses a new heuristic order crossover (HOC) and mutation (HSM) to keep the limited search domain.

The components required to develop HGA requires chromosome coding, initialization, genetic crossover operator, genetic mutation operator, and parent selection & termination rules.

Summary of HGA:
Step1: Chromosome Coding Scheme and Initialization: Chromosome Coding Scheme:
The complete chromosome of a candidate is divided into node fields equal to the number of nodes in a network. Refer Figure 2.

3	3	0	3	2	11

FIGURE 2: Part of Chromosomal structure of a candidate path

This structure uses the node indices and the distance weight between 2 nodes.
N i0 = Previous (Ni) N i1 = Ni
N i2 = dist (Ni)
Previous (Ni) is same as the predecessor array and dist (Ni) is same as an array of best estimates of shortest path to each vertex in the Dijkstra's algorithm.

Initialization: First node of every candidate path is the source node. So each chromosome (s,s,0). Other entries are random nodes, covers all other nodes in the graph.
Step2: Parent Selection, HOC, HSM and Termination: Parent Selection:
Here, algorithm chosen for the selection is Tournament Selection Algorithm. The idea behind this is to pick a pair at random, compare their fitness and the fittest is selected.
To find the fitness value we need to know the objective function (path cost). Path cost = sum of (dist (Ni)) where i = 1 to n.
Using this the fitness function value is calculated as: Fitness (Chromosome) = $\frac{1}{\text{Path Cost}}$

HOC:
Here node fields are chosen as the cut points. The portion of the first parent between them is copied to the offspring, the rest of the offspring is selected from the second parent with the following conditions:

1. The source node fields remains at the first position in new generated offspring.
2. While taking other nodes from the second parent, all nodes should be included only once in the offspring.

And the offsprings are evaluated and added as the new solution path candidate.
HSM:
Two node fields are chosen randomly and swapped given that the source node is never mutated. And again new mutated chromosome is evaluated. HOC and HSM don't generate new edges in any candidate path , they just adjust the initially generated nodes into a legal minimum cost path
Termination:
The algorithm can be terminated when the number of generation crosses an upper bound specified by the algorithm. With an increased number of generations, HGA converges to the optimal solution.

In year 2009, an algorithm based on Dijkstra for huge data [5] was proposed. In the paper author has pointed out the drawbacks of the Dijkstra's algorithm and proposed an algorithm as **adjacent node algorithm** which an optimization over Dijkstra's algorithm. He proved that his algorithm can save lot of memory and is more suitable for graph with huge nodes. The adjacent node algorithm makes improvement by improving the method of creating the adjacency matrix. First the number of the maximum adjacent nodes r is found. Then the adjacency matrix of n*r is made which is much smaller that n*n matrix. One more judgement matrix is made of order m*r. The shortest path is found with the help of both adjacency and judgement matrix. In their experiment, this algorithm performed 6 times better than the Dijkstra's algorithm for the data size of 12000 nodes.

In year 2009, a faster algorithm [2] has been proposed for SSSP problem. They have proposed an efficient method for implementing the Dijkstra's algorithm with the same assumptions. In addition to it two more assumption is made that : Let L= {l1,l2,........lk} be the set of distinct nonnegative edge weights given in an increasing order as part of the input stored as an array and the number of distinct edge lengths (k) is small. The author's solution is motivated by the "gossip" problem for social networks.

Two algorithms are proposed by the author in this paper:

1. Simple implementation of Dijkstra's algorithm that runs in O (m+nk) time.
2. Second algorithm is the modification of first algorithm by using binary heaps to speed up the FindMin() operation. Its running time is O(m log (nK/m)) if nK>2m.

Both the algorithms are identical to Dijkstra's algorithm. The difference is that it uses some additional data structures to carry out FindMin() operation. The algorithm is as follows:
Step1:

```
Function INITIALIZE()
S:={s}; T := V-{s}.
d(s):=0;  pred(s):=.Φ
for (each vertex vε T) do
    d(v)= ∞; pred(v)=. Φ
```

```
5: end for
6: for (t=1 to K) do
7:          Et(S):=. Φ.
8:          CurrentEdge(t):=NIL.
9: end for
10: for each edge(s, j) do
11:         Add(s, j) to the end of the list Et (S),where It =csj .
12:          if (CurrentEdge(t)=NIL) then
13:                  CurrentEdge(t):=(s, j)
14: end if
15: end for
16: for (t=1to K) do
17:         UPDATE(t)
18: end for
```

Step2:

```
Function NEW-DIJKSTRA()
1:  INITIALIZE ()
2:  while (T= Φ) do
3:          let r = argmin {f(t):1t K}.
4:          let (i, j) = CurrentEdge(r).
5:          d(j):=d(i) + Ir ; pred(j):=i.
6:          S= S Ụ{j}; T :=T - {j}.
7:          for (each edge (j,k) ε E(j)) do
 8. Add the edge (j,k) to the end of the list Et(S),where It =cjk .
 9:             if (CurrentEdge (t) = NIL) then
10:                  CurrentEdge (t): = (j,k)
11:             end if
12:         end for
13:         for (t = 1to K) do
14:                 UPDATE(t).
15:       end for
16: end while
```

Step3:

```
Function UPDATE(t)
1:Let (i, j) = CurrentEdge(t).
2: if (jεT) then
3:         f(t)=d(i)+cij
4:          return
5: end if
6: while ((j ε T) and (CurrentEdge(t).next != NIL)) do
7:        Let(i, j) = CurrentEdge(t).next.
8:         CurrentEdge(t)=(i, j).
9: end while
10: if (jT) then
11:        f(t)=d(i)+cij .
12: else
13:       Set CurrentEdge(t) to Φ.
14:       f(t)= ∞ .
15: end if
```

The initialization step takes O (n) time. The potential time taking operations are step 3 of New-Dijkstra and the Update procedure. In New-Dijkstra step3 takes O (k) per iteration of the while loop and O (nk) over all the iterations. Procedure Update is called O (nk) times and its total running time is O (m+nk). Iteration in which CurrentEdge (t) is not changed, running time is O (nk) and the iterations in which CurrentEdge(t) is changed, the running time is O(m).
So the total time taken by the algorithm is O (m+nk).

3. USEFULNESS OF ALGORITHMS IN VARIOUS APPLICATIONS

Thorup's algorithm [1] is much slower than the algorithms with the heaps as for the whole execution time is compared. It is very slow for SSSP due to the time of the construction of the data structures. Due to need of huge amount of memory and the complicated, large programs, Thorup's algorithm is not useful in practice today.

The algorithm in paper [2] works well for the graphs having smaller number of distinct edge lengths than the density of the graph.

The Heuristic Genetic algorithm [3] proved to be suitable for the network of different size and topology. HGA took reasonable CPU time to reach the exact solution and didn't variate much with increased input size.

The Augmented Shortest path algorithm [4] is suitable for the graphs with less value of f. According to author it is suitable for the road networks, electronic circuit designs etc.

The Adjacent Node algorithm in [5] is efficient for the huge data and takes less space. So it is suitable for the traffic analysis type of applications.

The algorithm based on partitioning of graph [6] although performed better than Dijkstra's algorithm for some cases but for large networks its performance is degraded than that of Dijkstra's.

4. CONCLUSIONS

In this paper it is tried to be explained- first, what are the different algorithms for the SSSP problem. Second, how do they perform in comparison of the Dijkstra's algorithm. Third, which algorithm is suitable for a particular application or situation.

5. REFERENCES

[1] Y. Asano, H. Imai, " Practical Efficiency of the Linear Time Algorithm for the Single Source Shortest Path problem", *Journal of the Operations Research, Society of Japan, Vol. 43, No. 4; 2000.*

[2] J. B. Orlin, K. Madduri, K. Subramani, M. Williamson, "A faster algorithm for the Single Source Shortest Path problem with few distinct positive lengths"*, Journal of Discrete Algorithm; 2009.*

[3] B. S. Hasan, M. A. Khamees, A. S. H. Mahmoud, "A Heuristic Genetic Algorithm for the Single Source Shortest Path Problem", *IEEE International Conference on Computer Systems and Applications; 2007.*

[4] P. P. Mitra, R. Hasan M. Kaykobad, "On Linear time algoritm for SSSP Problem", *ICCIT, 2000.*

[5] Zhang Fuhao, L. Jiping, "An algorithm of shortest path based on Dijkstra for huge data", *6th International Conference on Fuzzy Systems and Knowledge discovery, 2009.*

[6] R. H. Mohring and H. Schilling, "Partitioning Graphs to Speedup Dijkstra's Algorithm", *ACM Journal of Experimental Algorithmics, Vol. 11, Article No. 2.8, Pages 1-29, 2006.*

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Computer Science and Security (IJCSS)* is a refereed online journal which is a forum for publication of current research in computer science and computer security technologies. It considers any material dealing primarily with the technological aspects of computer science and computer security. The journal is targeted to be read by academics, scholars, advanced students, practitioners, and those seeking an update on current experience and future prospects in relation to all aspects computer science in general but specific to computer security themes. Subjects covered include: access control, computer security, cryptography, communications and data security, databases, electronic commerce, multimedia, bioinformatics, signal processing and image processing etc.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 6, 2012, IJCSS appears in more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCSS LIST OF TOPICS

The realm of International Journal of Computer Science and Security (IJCSS) extends, but not limited, to the following:

- Authentication and authorization models
- Computer Engineering
- Computer Networks
- Cryptography
- Databases
- Image processing
- Operating systems
- Programming languages
- Signal processing
- Theory
- Communications and data security
- Bioinformatics
- Computer graphics
- Computer security
- Data mining
- Electronic commerce
- Object Orientation
- Parallel and distributed processing
- Robotics
- Software engineering

CALL FOR PAPERS

Volume: 6 - **Issue:** 6 – December 2012

i. Paper Submission: September 30, 2012 **ii. Author Notification:** November 15, 2012

iii. Issue Publication: December 2012

CONTACT INFORMATION

Computer Science Journals Sdn BhD
B-5-8 Plaza Mont Kiara, Mont Kiara
50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2012
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org