IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 5 Issue 11

November 2014

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the November 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 11 November 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Handwritten Pattern Recognition Using Kohonen Neural Network Based on Pixel Character

Lulu C. Munggaran, Suryarini Widodo, Cipta A.M
Faculty of Computer Science & Inf.Technology
Gunadarma University
Depok, Indonesia

Nuryuliani
Faculty of Industrial Technology
Gunadarma University
Depok, Indonesia

***Abstract*—Handwriting is the human way in communicating each other using written media. By the advancement in technology and development of science, there are a lot of changes of technology in terms of communication with computer through handwriting. Therefore, it is needed computer able to receive input in the form of handwriting data and able to recognize the handwriting input. Therefore, this research focuses on handwritten character recognition using Kohonen neural network. The purpose of this research is to find handwriting recognition algorithm which can receive handwriting input and recognize handwritten character directly inputted in computer using Kohonen neural network. This method studies the distribution of a set of patterns without any class information. The basic idea of this technique is understood from how human brain stores images/patterns that have been recognized through eyes, and then able to reveal the images/patterns back. This research has been successful in developing an application to recognize handwritten characters using Kohonen neural network method, and it has been tested. The application is personal computer based and using a canvas as input media. The recognition process consist of 3 stages layer: Input layer, Training Layer and Hidden Layer. The Kohonen neural network method on handwritten character recognition application has good similarity level of character patterns in character mapping process.**

Keywords—handwriting; recognition; Kohonen neural network; similarity; character

I. Introduction

Handwriting is the human way in communicating using written media. Nowadays, there are a lot of changes of technology in terms of communication. Handwriting offers an attractive and efficient method to interact with computer, such as a tool developed at this time that is able to receive input in the form of handwriting data. The tool also requires a method to recognize input in the form of handwriting data. Research on handwriting recognition issue becomes important to provide solutions of the problems above.

There is difference of writing patterns for each person's writing so that the way the computers recognize writing patterns should be noted. Writing pattern recognition [1] is intended for computer to be able to recognize the letters/characters by converting images, either printed or handwritten, into codes [1]. Handwriting pattern recognition can be done online or offline [2]. There are many methods that can be used to identify patterns of writing, one of method is neural network [3].There are many methods that can be used to recognize writing patterns such as template matching, character decomposition, wavelets, neural network, etc.

Ralph Niels and Louis VUURPIJL, did a research of handwritten character recognition using template matching method. This method's advantage is that this method produces a higher level of recognition. And this method's disadvantage is that users should be able to specify templates to compare characters to be tested [4]. Template matching method was also performed by Jakob Sternby that results in the recognition rate of 82.4%. The method developed by Sternby is suitable for handwriting recognition based on graph segmentation system but it requires a long processing time because it should match input characters with all templates [5]. Recognition by decomposing character was conducted by E. Gómez Sánchez, et al. Their method's advantage is that it is suitable for real-time processing, consistently, for the combination of curvature multiple forms particularly on character segmentation of upper case characters, however, this method depends on how the character is written, and handwriting which is written very fast provides troubles. The recognition rate of digit is 82.52%, for uppercase character is 76.39% and for lowercase character recognition is 58.92% [6].

Neural network has an ability to learn information already received. Based on this method, neural network is divided into two, namely supervised learning and unsupervised learning. The difference of the two methods is in terms of target output. In supervised learning there is target output while unsupervised learning has no target output. From both methods, the most appropriate method to recognize patterns is unsupervised learning. The example of unsupervised learning is Kohonen network method.

Kohonen Neural Network method studies the distribution of a set of patterns without any class information. The basic idea of this technique is understood from how human brain stores images/patterns that have been recognized through eyes, then human brain is able to reveal the images/ patterns back. Therefore, this model is widely used in object recognition or visual images. There are a lot of interesting problems in pattern recognition; some of them are fingerprint pattern recognition, facial pattern recognition, handwritten character pattern recognition and so on.

Based on the pattern recognition problems, it has been developed handwritten character pattern recognition application with neural network [7]. The handwritten character pattern recognition is aimed for computer in order to be able to recognize handwritten character by converting images of characters, either printed or handwritten into digital data. The data are saved into data file so the data can be edited and manipulated on a computer.

To solve handwriting pattern recognition problems in neural network, it is used kohonen algorithms; neural network is analysed in terms of the accuracy of handwriting recognition on patterns and time required for training of handwriting patterns.

This research is aimed to recognize handwritten characters from characters “a-z”, uppercase and lowercase characters using Kohonen Neural Network in order to allow users to recognize handwritten character patterns of character drawing process

II. Literature Review

A. Kohonen Neural Network

Kohonen Neural Network method is an unsupervised learning process studying distribution of a set of patterns without any class information. The basic idea of this technique is understood from how human brain stores images/patterns that have been recognized through eyes, and then human brain is able to reveal the images/ patterns back. Therefore, the application of this model is widely used in object recognition or visual image. A brief description of the form of a series of kohonen network can be seen in Fig. 1.

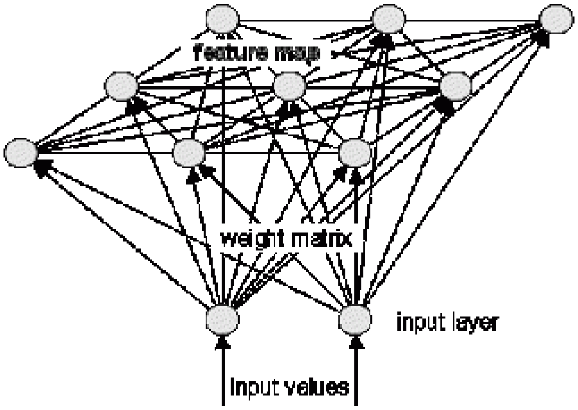


Fig. 1. KohonenNeural Network [8]

There are several studies on the use of Kohonen Neural Networks in Handwriting. The study on Online Recognition of handwritten Arabic Characters Using a Kohonen Neural Network was conducted by Mezghani, et al [9]. This study used 18 arabic alphabet character forms, with 7400 samples. The kohonen neural network used was Kohonen memory (known as Kohonen Self Organizing feature map) used to represent all points of a source into a smaller number of points in the target space. The memory was organized as 2 dimensional arrays. This study resulted in 88.38% recognition rate on writing characters without limitation of writers.

Another study using Kohonen Neural Network was Vishwaas, et al [10]. Vishwass combined Direction Based Stroke Density (DSD) with Kohonen Neural Network for Kannada Writing Recognition. The study used 49 characters and 10 numbers of Kannada (Kannada Language) written by 20 different people. This study resulted in 94.4% recognition accuracy.

B. Handwriting Recognition

Handwriting recognition is an ability of a computer to receive input in the form of understandable handwriting. Principally, handwriting recognition requires optical character recognition. The handwriting recognition system completely handles formatting, performs segmentation, and finds the most appropriate word. Handwriting recognition is used for editing, annotation, other applications that are difficult to interact directly, and applications using direct manipulation and pointing. Tablet is a powerful tool as the input of handwriting because tablet can receive both text and graphics. There are many electronic devices that can be used as a tool in Handwriting recognition such as a mobile device/smartphone [11,12,13]. Research on the utilization of input devices such as electronic pen was also conducted by Sreeraj [14].

Handwriting consists of a time sequence of stroke (movement), i.e. movement of writing started from pen down to pen up, character writing is usually formed in a sequence, one character is complete before starting the next one, a character usually starts from left to right. In some cases there are exceptions. For example in English script, crosses (for t and x) and point (for i and j) tend to be a delay in writing.

III. Proposed Method

The method proposed in this research was handwritten character recognition written on canvas using Kohonen neural network. Canvas is an area to write input characters. The application is operated on a personal computer. The application is built using Java programming language. The research stages proposed are outlined in Fig. 2 as follows.

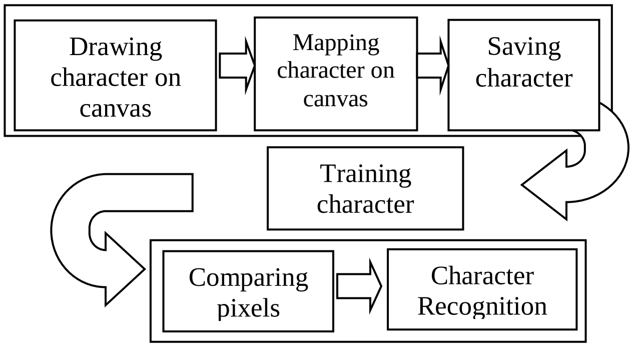


Fig. 2. Stages of Handwritten Character Recognition

In general, there are six steps of handwritten character recognition process. : The first step is drawing character on canvas; the 2nd step is mapping image of character; the 3rd steps is saving image of character data; the 4th step: training data; the 5th steps is matching data by comparing the results of mapping images with character data which have been saved at the preceding step, finally the last step is recognition: the characters can be recognized or not. The steps could be divided in the 3 stages layer: Input layer, Training Layer and hidden layer.

In the first stage is Input layer. Drawing character on canvas process becomes the input layer process. Input layer is very important to enter information into other processes. In this process users should draw a character to continue the next process. The drawing area can be seen in fig 3.

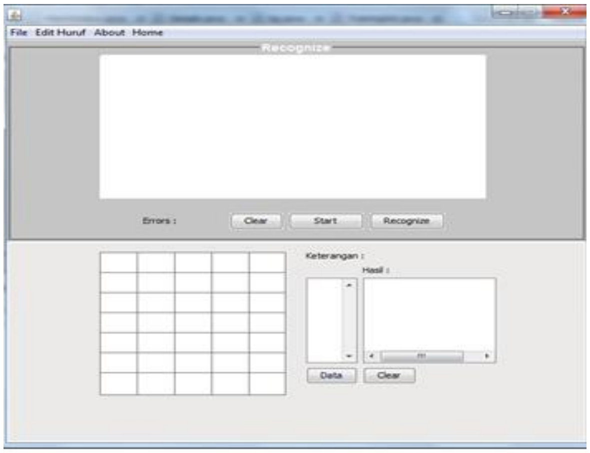


Fig. 3. The drawing area

This process is the main layer which does drawing and the core of subsequent processes. This layer will scan character from image. Each image formed is a collection of x-coordinate and y-coordinate points. When mouse touches canvas, the initial value of x-coordinates and y-coordinates will be saved in the memory, and when mouse is released the initial values of x-coordinates and y-coordinates are saved. Between initial value and final value of coordinates there will be whole-values of x and y coordinates saved in memory when mouse is pressed (Mouse Pressed) and mouse is dragged (Mouse Dragging). Clipping process is performed when mouse is released from dragging process; image will be framed by a box along with parsing of coordinate values to mapping process and data recognition. The results of clipping image process save coordinate value result from process of mouse pressed and mouse dragging. To frame image with a box it has to take the largest and smallest values of x coordinates and y coordinates. The program will form a straight line connected between the largest value and the smallest value.

The next step is mapping characters, i.e. the taking process of image cutting results in which its values are adjusted to the resolution of existing map size. The map size is made with the resolution of 7x5 pixels.

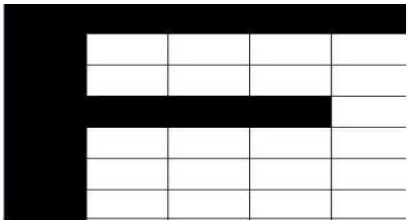


Fig. 4. Mapping of Character “F” Image

For example, in Fig. 4 character “F” has been already processed in mapping process, its value is F={1,1,1,1,1,1,0,0,0,0,1,0,0,0,0,1,1,1,1,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0}.

This value is obtained from the results of 7x5 pixels of pixel box in which the allocated value has 1 value and the unallocated has 0 value. The value is calculated based on columns from top to bottom.

Next, the value of mapping process is saved in a text file with .dat extension. Thirty-five pixels value of each character saved in this process.

The second stage is training process (training layer). Training layer process will train and scan character depicted in input layer process. This scanning process will read and save image pixels into database. The information read by this layer will be saved in hidden layer.

The third stage is the process where characters are entered, can also be saved and taken whenever required while information saved is required, the character taking happens from hidden layer process. The character scanned by training layer can be saved in this layer. This layer is not visible by users, so that when character is drawn, the character will be scanned in this layer and if the image character matches to the existing data, the result will be displayed. In this hidden layer process character will be taken and saved. All information will be taken and applied for further proceedings.

The final step is character recognition and selected character approach process (matching and output). Matching and output is the final process where the matching operation is done. If there is image formed in drawing process, training layer will scan the image and character matching is done, after character matching is finished, the output result will be displayed to users. A procedure to display the scanning result will be conducted on layer output with a clear and efficient way.

IV. Result

The data used in this research are characters a-z and characters A-Z written on canvas. The size of handwritten character data is 7x5 pixels. The recognition process of some handwritten characters as follow:

1) Drawing Characters

The first step of handwritten character recognition process is drawing characters on canvas. The character drawing process canseen in Fig. 5, when mouse is released from canvas, there will be formed a straight line at the top, right, left, and bottom of image then the image is framed in a box.

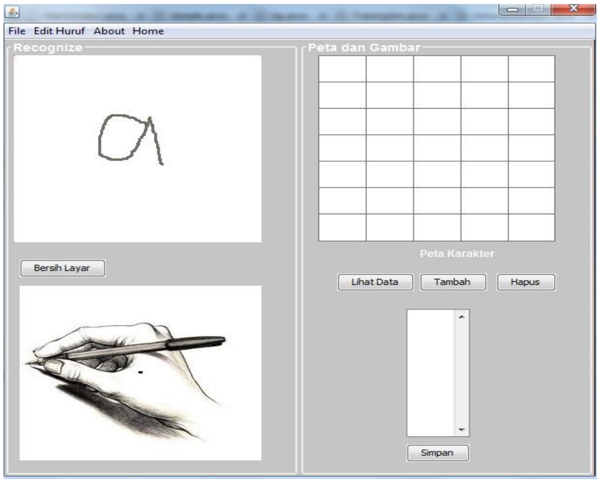


Fig. 5. Drawing Character

2) *Mapping Image*

When mouse is released, there will be performed mapping image process into canvas map; the result can be seen in Fig. 6.

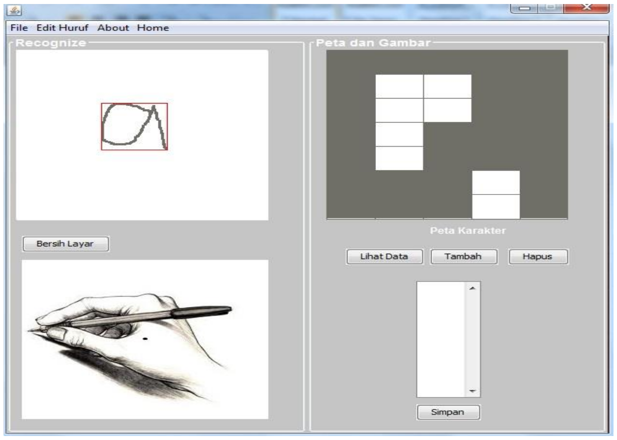
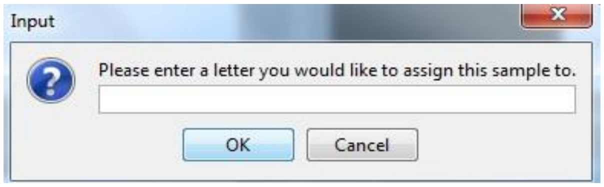


Fig. 6. Mapping Image

3) *Saving Character Data*

To add character data, it can be performed by pressing plus sign button, and there will be a form of character input, the result will be shown in Fig. 7.

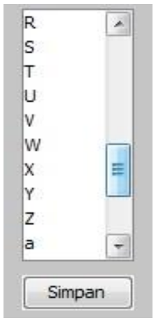


Input

Please enter a letter you would like to assign this sample to.

Fig. 7. Character Fill-In Form

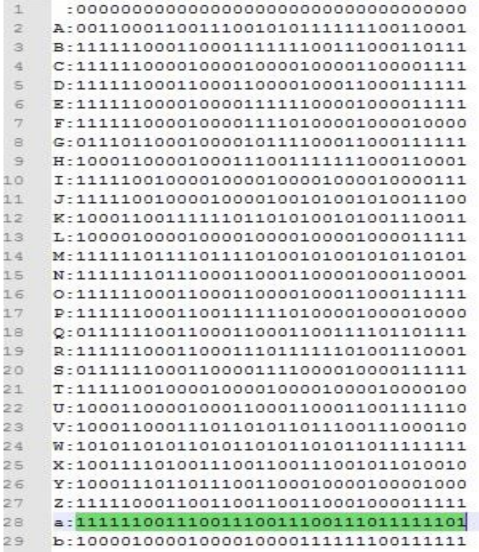
After character input form is filled, the input character will appear on a list of data as in Fig. 8, to save the list of data into sample.dat file, it can be done by pressing save button.



R
S
T
U
V
W
X
Y
Z
a

Fig. 8. Character Data in List of Data

The result of character input can be seen in sample.dat and the result can be seen in Fig. 9. The values saved in sample.dat file are in the form of characters of data and binary number of 0 and 1



```
 :00000000000000000000000000000000000
A:00110001100111001010111111100110001
B:11111100011000111111100111000110111
C:11111100001000010000100001100001111
D:11111100011000110000100011000111111
E:11111100001000011111000010000011111
F:11111100001000011110100001000010000
G:01110110001000010111100011000111111
H:10001100001000111001111111000110001
I:11111001000010000100001000010000111
J:11111001000010000100101001010011100
K:10001100111111011010100101001110011
L:10000100001000010000100001000011111
M:11111101111011110100101001010110101
N:11111110111000110001100001000110001
O:11111100011000110000100011000111111
P:11111100011001111110100001000010000
Q:01111110011000110001100111101101111
R:11111100011000111011111101001110001
S:01111110001100001111000010000111111
T:11111001000010000100001000010000100
U:10001100001000110001100011001111110
V:10001100011101101011011100111000110
W:10101101011010110101101011011111111
X:10011110100111001100111001011010010
Y:10001110110111001100010000100001000
Z:11111000110011001100110001000011111
a:11111100111001110011100111011111101
b:10000100001000010000111111100111111
```

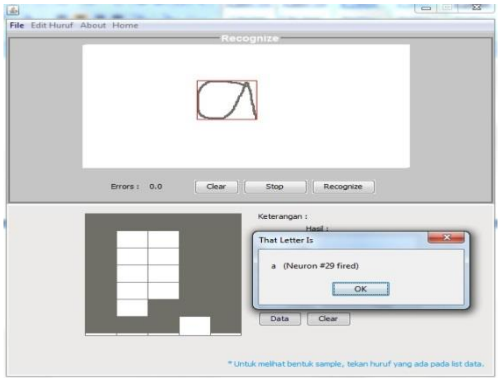
Fig. 9. Data saved in sample.dat

4) *Comparing Pixels*

Pixel compared is the image pixel of handwritten character input result on canvas with the data saved previously in sample.dat file.

5) *Character Recognition*

Pixel comparison process is performed in order to be able to recognize characters. After that, a testing is done to recognize handwritten characters as shown in Fig. 10 and Fig.11.



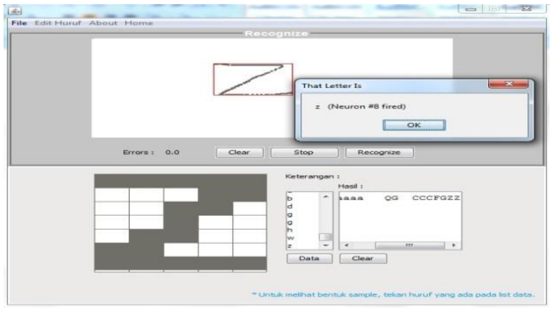
That Letter Is

a (Neuron #29 fired)

OK

* Untuk melihat bentuk sample, tekan huruf yang ada pada list data.

Fig. 10. Recognition of Character “a”



That Letter Is

z (Neuron #8 fired)

OK

b
d
g
g
h
w
z

aaaa QG CCCFGZZ

* Untuk melihat bentuk sample, tekan huruf yang ada pada list data.

Fig. 11. Recognition of Character “z”

The complete results of testing done to users who write characters “a to z” as many as 10 times can be seen in Table 1:

TABLE I. CHARACTER “A-Z” RECOGNITION TESTING RESULTS

Character Input	Occurrences of characters									
Persons	1	2	3	4	5	6	7	8	9	10
A	A	A	A	A	A	A	A	A	A	A
B	█	█	█	B	█	B	B	█	█	█
C	C	C	C	█	C	C	C	C	C	C
D	█	█	█	█	█	█	█	█	█	█
E	E	E	E	E	E	E	█	E	E	E
F	█	█	█	F	F	F	█	F	F	F
G	G	G	B	█	G	G	G	█	G	█
H	█	█	█	█	█	H	H	H	H	H
I	I	█	I	I	█	I	I	I	I	I
J	J	J	█	J	█	J	J	J	J	J
K	K	█	█	K	█	█	█	█	█	K
L	L	L	L	L	L	L	L	L	L	L
M	M	M	M	M	M	M	M	M	█	M
N	N	N	N	N	N	N	N	N	N	N
O	O	O	O	O	O	O	O	O	O	O
P	P	█	P	P	P	█	█	P	█	P
Q	Q	Q	█	█	Q	█	Q	█	Q	█
R	R	R	R	█	█	R	█	█	█	█
S	S	S	S	S	█	S	S	S	S	█
T	T	T	T	T	T	T	█	T	T	T
U	U	U	U	U	U	U	U	U	U	U
V	V	V	V	V	V	V	V	V	V	V
W	W	W	W	W	W	W	W	W	W	W
X	X	█	X	X	Y	█	█	X	█	█
Y	Y	█	Y	Y	Y	█	Y	█	█	Y
Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
Number of matches	22	17	19	20	18	20	18	19	19	19

From the testing results in the table above it shows that the recognizable characters are character “a, l, n, o, u, v, w and z”. The characters can be recognized 100%. Whereas, the least recognizable character is “d”, character d is more recognizable as character “o” and “q”. The unrecognizable writing of character “d”.

V. CONCLUSION

Handwritten character recognition using Kohonen neural network can recognize characters based on handwriting pattern approach with the character data available. The recognition process consist of 3 stages layer: Input layer, Training Layer and Hidden Layer. In Input layer: user draw a character. On the testing, users asked to write characters “a, l, n, o, u, v, w to z “ on canvas area as many as 10 times. The size of handwritten character data is 7x5 pixels. The data read and save in Training Layer. The recogniton result shown in the hidden layer. The experiment shows that character “a” is the most recognizable character and the most unrecognizable character is character “d”. The handwritten character recognition with this method has similarity character pattern with mapping character process of 87.5%. The good results of handwriting recognition can be obtained by drawing process of character writing and can be adapted to the character data which saved in sample.dat file. The results will not be appropriate if the drawing process producing mapped images has inappropriate pixel values to the existing pixel values in sample.dat file. The contribution of this research is that it can be used to introduce the writing learning for early age children.

ACKNOWLEDGEMENT

For the Directorate General of Higher Education, Ministry of Education and Culture of Republic of Indonesia and Gunadarma University, thank you for supporting this research.

REFERENCES

[1] Asworo, Ed, “Compatration Between Kohonen Neural Network and Learning Vector Quantization Methods on Real Time Handwritting Recognition System”, Institute of Technology Surabaya.

[2] Plamondon, Rejean and Sargur N Srihari, “Online and Off-Line Handwriting Recognition: A Comprehensive Survey”, IEEE Transactions on (Volume:22 , Issue: 1 , Jan 2010.

[3] Senior, A.W.,and Robinson, A.J., An off-line cursive handwriting recognition system, IEEE Transactions on Pattern Analysis and Machine Intelligence, (Volume:20 , Issue: 3) Mar 1998, P:309 - 321, ISSN : 0162-8828

[4] Niels, Ralph and Vuurpijl Louis. (2005), Using Dynamic Time Warping for Intuitive Handwriting Recognition, Advances in Graphonomics: Proceedings of IGS 2005

[5] Sternby, Jakob. (2005), Structurally Based Template emplate Matching of On-line Hand Handwritten written Characters

[6] E. Gómez Sánchez., et al, “On-Line Character Analysis and Recognition with Fuzzy Neural Networks”,Intelligent Automation and Soft Computing, vol. 7, No. 3, pp. 161-162, 1998.

[7] E. Anquetil and H. Bouchereau, “Integration of an on-line handwriting recognition system in a smart phone device,” in Pattern Recognition,

2002. Proceedings.16th International Conference on, vol. 3. IEEE, 2002, pp. 192–195.

[8] Jochen Fröhlich, "Neural Networks with Java: Neural Net Components in an Object Oriented Class Structure", Fachhochschule Regensburg , Department of Computer Science, 1997, url address: http://fbim.fh-regensburg.de/~saj39122/jfroehl/diplom/e-idex.html

[9] Mezghani, Neila., Mitiche, Amar., and Cheriet Mohamed, "On-Line Recognition of Handwritten Arabic Characters Using A Kohonen Neural Network", Proceeding of The Eigtht International Workshop of Frontiers in Handwriting Recognition, IEEE, 2002.

[10] Vishwaas, M., Arjun, N.M., and Dinesh, R.,"Handwritten Kannada Character Recognition Based on Kohonen Neural Network", International Conference on Recent Advances in Computing and Software Systems from 25-27th April, IEEE, 2012

[11] G. Robert and Jones, "Emerging Technologies Mobile Computing Trends : Lighter, Faster and Smarter", Vol. 12, No. 3, October 2008.

[12] Sanchez, R. M. Sachez, and P. Garcia, Eds,"On-line Character Analysis and Recognition with Fuzzy Neural Network", 1998

[13] Sreeraj, M and Idicula, S. M, "On-line Handwritten character recognition using kohonen networks", Nature& Biologically Inspired Computing, 2009. NaBIC 2009. Des 2009 Source: IEEE Xplore

[14] Walid A.Salameh and Mohammed A.Otair, Online Handwritten Character Recognition using an Optical Backpropagation Neural Network, Issues in Informing Science and Information Technology, pp.787-795.

Basic Study for New Assistive Technology Based on Brain Activity during Car Driving

Hiroaki Inoue
Tokyo University of Science, Suwa
Research course of Engineering/Management
Chino-city, Japan

Fumikazu Miwakeichi
The Institute of Statistical Mathematics
Spatial and Time Series Modeling Group
Tachikawa-city, Japan

Shunji Shimizu, Noboru Takahashi and Yasuhito Yoshizawa
Tokyo University of Science, Suwa
Department of Computer Engineering
Chino-city, Japan

Nobuhide Hirai
Tokyo Medical and Dental University
Health Service Center
Bunkyou-ku, Japan

Hiroyuki Nara
Hokkaido University
Graduate School of Information Science and Technology
Sapporo-city, Japan

Senichiro Kikuchi, and Satoshi Kato
Jichi Medical University
Department of Psychiatry
Shimotsuke-city, Japan

Eiju Watanabe
Jichi Medical University
Department of Neurosurgery
Shimotsuke-city, Japan

***Abstract*—Recently, it is necessary to develop a new system which assists driving car and wheelchair as aged society. The final our purpose in this research is to contribute to developing of assistive robot and related-apparatus. In terms of developing a new system, we thought that it is important to examine behaviors as well as spatial recognition. Therefore, experiments have been performed for an examination of human spatial perceptions, especially right and left recognition, during car driving by using NIRS. In previous research, it has been documented that there were significant differences at dorsolateral prefrontal cortex at left hemisphere during virtual driving task and actual driving. In this paper, we performed measuring the brain activity during car driving by using NIRS. And we performed statistical analysis of the brain activity. The purpose of this paper is discovering the brain region which was involved in decision making when human drive a car and considering between human movement and brain activity during car driving.**

***Keywords*—brain information processing during driving task; spatial cognitive task; determining direction; NIRS**

I. Introduction

Human movements change relative to their environment. Nevertheless, human recognizes the new location and decides what behavior to take when they move by a vehicle. It is important to analyze the human spatial perception for developing autonomous robots or automatic driving. In the previous study, the relation of the theta brain waves to the human spatial perception was discussed in [1][2]. When humans perceive space, for example, try to decide the next action in a maze, the theta brain waves saliently appear. This means human have a searching behavior to find a goal at an unknown maze. From the side of human navigation, Maguire et al. measured the brain activations using complex virtual reality town [3]. But, every task is notional and the particulars about the mechanism that enables humans to perceive space and direction are yet unknown. Also, Brain activities concerned with cognitive tasks during car driving have been examined. For example, there was a report about brain activity when disturbances were given to subjects who manipulated a driving simulator. Also, power spectrums increased in beta and theta bands [4]. However, there is little report on the relationship among right and left perception and driving task.

So, we performed experiments in which perception tasks were required during virtual car driving using Near Infrared Spectroscopy (NIRS) [5]. From experimental results, there were significant differences at dorsolateral prefrontal cortex in left hemisphere via one-sample t-test when subjects watched driving movie and moving their hand in circles as if handling a steering wheel [6].

In addition, we performed experiments in real-space, which were performed by taking NIRS in the car, and measured the brain activity during actual driving. A purpose in this experiment was to measure and analyze the brain activity during actual driving to compare results between virtual and actual results. As a result, there were significant differences at similar regions [7][8]. In addition, we measured the brain activity of frontal lobe, which is related to behavioral decision-making, during car driving in different experimental design from previous one to verify previous results [9][10].

It is well known that higher order processing, such as memory, judgment, reasoning, etc. is done in the frontal lobe [11]. Human recognize the various information of car surrounding situation and perform car driving movement. We tried to elucidate the mechanism of information processing of the brain by analyzing data about human brain activity during car driving. Measuring of the brain activity by using NIRS is always affected by human body movement. In this paper, we tried to measure the brain activity during car driving and perform detailed analysis about human movement of car driving. Also, the goal of this study is to find a way to apply this result to new assist system.

II. Experiment

A. *Brain activity on virtual driving*

1) Measuring the Brain activity when subjects saw the driving movie.

We performed measurement which was under the virtual environment before performing measuring of actual car driving. And we confirmed the possibility of brain activity measurement. Experimental method was as follows.

The movie is included two scenes at a T-junction in which it must be decided either to turn to the right or left. In the second scene, there is a road sign with directions. We used nine kinds of movies in about one minute(Fig.1). Before showing the movie, subjects were given directions to turn to the right or left at the first T-junction. They were also taught the place which was on the road sign at the second T-Junction. They had to decide the direction when they looked at the road sign. They were asked to push a button when they realized the direction in which they were to turn.

The subjects for this experiment were eight males who were right handed. They were asked to read and sign an informed consent regarding the experiment.

An NIRS (Hitachi Medical Corp ETG-100) with 24 channels (sampling frequency 10 Hz) was used to record the density of oxygenated hemoglobin (oxy hemoglobin) and deoxygenated hemoglobin (de-oxy hemoglobin) in the frontal cortex area.

2) Measurements of the brain activity when subjects performed the movement which was imitated the steering.

In this experiment, measuring was performed by NIRS, made by SHIMADZU Co. Ltd with 44ch. Five subjects were healthy males in their 20s, right handed with a good driving history. They were asked to read and sign an informed consent regarding the experiment.

Fig.1. Perspective of the video which was used in the experiment

The subject was asked to perform simulated car driving, moving their hand in circles as if using a steering wheel. A PC mouse on the table was used to simulate handling a wheel, and NIRS (near-infrared spectroscopy) to monitor oxygen content change in the subjects' brain. NIRS irradiation was performed to measure brain activities when the subject sitting on a chair make a drawing circle line of the right or left hand. Also, we set the experimental condition which was the direction of clockwise or counterclockwise. The part of measurement was the frontal lobe. The subject was asked to draw on the table a circle 30 cm in diameter five times consecutively, spending four seconds per a circle. The time design was rest (10 seconds at least) – task (20 seconds) – rest (10 seconds) - close rest.

B. *Brain activity during actual car driving*

1) Measuring the brain activity when subjects drove on actual car

In general roads, experiments were performed by taking NIRS in the car, and measuring the brain activity when car driven by subjects was went through two different intersections. Six subjects were a healthy male in their 20s, right handed with a good driving history. They were asked to read and sign an informed consent regarding the experiment. In all experiments, measuring was performed by f-NIRS (Functional Near Infrared Spectroscopy), made by SHIMADZU Co. Ltd [11].

Subjects took a rest during 10 seconds at least with their eye close before driving task and they drove a car during about 600 seconds. Finally, subject closed their eyes for 10 seconds again after task. Then, the brain activity was recorded from the first eyes-closed rest to the last eyes. Subjects were given directions to turn to the right or left at the first T-junction during driving task. They were also taught the place which was on the road sign at the second T-junction. And, they were given the place where they have to go to. So, they had to decide the direction when they looked at the road sign.

A trigger pulse was emitted on stop lines at T-Junctions to use as a measuring stick for the analysis. Also, we recorded movie during the experiment from a car with a video camera aimed toward the direction of movement (Figure. 1). Recorded movies were used to exempt measurement result including disturbances, such as foot passengers and oncoming cars, from analysis. Figure.2 and Figure.3 shows one sample of T-junction.

2) Verification Experiment

To conduct verification for experimental results in previous experiment, we performed additional experiment which was achieved in a similar way. In this experiment, experimental course was different from previous one. While previous one was included two T- junctions in which there was road sign at second one and not at first one per a measurement, there were multiple T-junctions.

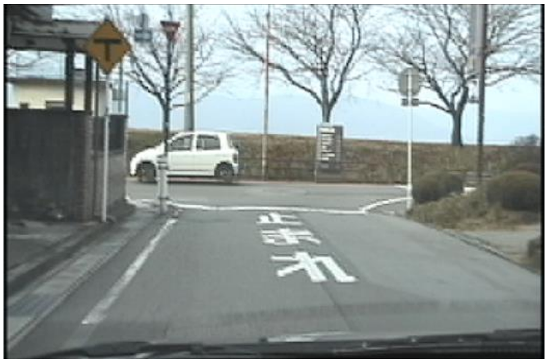


Fig.2. Sample of first T-Junction

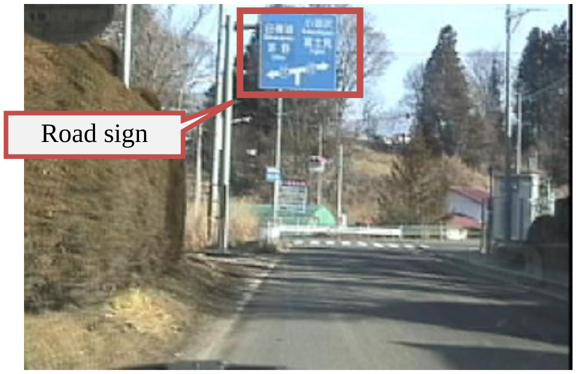


Fig.3. Sample of second T-Junction

Scenes of turning at the T-junction were recorded on the movie. There were nine scenes in movies. In five scenes, the road sign was contained. In four scenes, the road sign was not contained.

Subjects were twelve males who were all right-handed. They drove a car during about 20 minutes after a rest during 10 seconds at least with their eyes close. Subjects were enlightened about turning direction and the place on which road signs was at T-junction during measurement. And, they arbitrarily decided the direction to turn when they confirmed road signs. Also, a trigger pulse was emitted in the same way.

3) Detailed analysis based on driving behavior

In this analysis, we focused on the movement of subjects during car driving. We verified that there are no relationship between decision-making of the direction and movement. We have attention to the movement of subjects arms and ocular. In previous research we performed, stop line at T-junction was used as a trigger. But, brain activity in T-junction involved movement task such as turning steering wheel, changing neck direction, hitting the brake. So, it is thought that brain activity derived from cognitive tasks was overwritten with brain activity due to movement tasks. Therefore, we tried to analyze brain activity on the basis of ocular motions to examine significant differences with cognitive tasks.

III. Experimental Results

A. Brain activity on virtual driving

1) Measuring the Brain activity when subjects saw the driving movie.

On the whole, the variation in de-oxy hemoglobin was smaller than in the oxy hemoglobin. It is considered that increasing oxy hemoglobin was caused by brain activity. Also, there was a great increase in channel 18(around #10 area of the dorsolateral prefrontal cortex of the right hemisphere). This might be the variation based on the spatial perceptions.

Next, differences were investigated concerning the subject's brain activity. As the first case, it was when the vision was directed after having been told the direction. As the Second, it was when the vision was directed after having been decided the direction under the road sign. We were trying to find a brain region involved in decision making that determines the direction by comparing these brain activity data in statistical analysis[12]. The experiment has been performed under the same conditions in any task without difference in instruction and difference between left and right. Therefore, the difference in instructions and difference between left and right and has appeared as a test result.

2) Measurements of the brain activity when subjects performed the movement which was imitated the steering.

In previous experiment, significant difference appeared due to the difference in the direction to turn left and right. In this experiment, we use PC mouse because, we confirm this results were due to the difference of the left and right rather than by the operation of the steering wheel. During the motion, the increase of oxy hemoglobin density of the brain was found in all subjects. The different regions of the brain were observed to be active, depending on the individual. The subjects were to be observed 1) on starting, and 2) 3-5 seconds after starting moving their 3) right hand 4) left hand 5)clockwise 6)counterclockwise.

We perform one sample t-test to analysis. Sample of t-test were oxy–hemoglobin at starting task and average of the average value of oxy-hemoglobin from 3 second to 5 second seconds after the start of the task.

Although some individual variation existed, the result showed the significant differences and some characteristic patterns. This pattern was increasing oxy hemoglobin density on starting task and decreasing oxy hemoglobin after end of task. The obtained patterns are shown as follows. Regardless of 1), 2), 3) and 4) above, the change in the oxy hemoglobin density of the brain was seen within the significant difference level 5% or less in the three individuals out of all five subjects. The part was the adjacent part both of left pre-motor area and of left prefrontal cortex. Especially, in the adjacent part of prefrontal cortex a number of significant differences were seen among in four out of five subjects. Next more emphasis was put on the rotation direction: 5) clockwise or 6) counterclockwise. No large density change was found in the brain with all the subjects employing 6). Figure.4 shown the brain regions which were observed the statistical significant difference in each sample data.

In the Figure.4, brain regions which were draw by red. It is well known that in the outside prefrontal cortex higher order processing is done such as of behavior control. It is inferred that the pre- motor area was activated when the subjects moved the hand in the way stated above because the pre-motor area is responsible for behavior control, for transforming visual information, and for generating neural impulses controlling.

Subjects performed experiments in random order experimental conditions. Therefore it is not able to consider that experimental time and sequence effect of experimental conditions.

B. Brain activity during actual car driving

1) Measuring the brain activity when subjects drove on actual car

At the first, Hb-oxy was increased in overall frontal lobe after start of operation. This tendency was common among subjects. After that, Hb-oxy was decreased as subjects adjusted to driving the car. This meant that the brain activity changed from collective to local activities.

We performed one sample t-test as similar as previous virtual experiment. Fig. 5 and Fig.6 is shown the brain region where the statistical significant difference was seen.

2) Verification Experiment

Various tendencies among individuals were observed in comparison with results in result of actual car driving.

However, there were tendency that oxy-Hb was increased when car turned left or right at T-junctions and oxy-Hb was decreased during going straight.

Analysis method was the same as previous one. Though Gaps were shown regions at which there were significant differences, there were significant differences in # 46 and pre-motor regions which was surrounded by red circles, too (Fig. 5 and Fig.6). In the analysis, measurement results including disturbance at T-junctions were excluded as analysis object. The statistical significant difference was seen in the region like the previous experiments.

3) Detailed analysis based on driving behavior

The analysis was performed one-sample t-test within the significant difference level 5% or less between brain activity before and after looking at road sign. Each of sample data with respect to each direction which subject had to go at next T-junction. As a consequence of analysis, there were significant differences at interior front gyrus of frontal lobe of left hemisphere without reference of direction (Fig. 8 and Fig. 9).

Also, we paid the attention to the movement of subjects for driving and performed one-sample t-test on the brain activity data. There are three analysis points which are the change of eyes vertical movement when subjects looked to road sign, the change of neck horizontal movement when subjects seen around situation and the change of hand movement when subjects turn a steering wheel. Figure was shown the collection method of the brain activity data to use for one sample t-test.

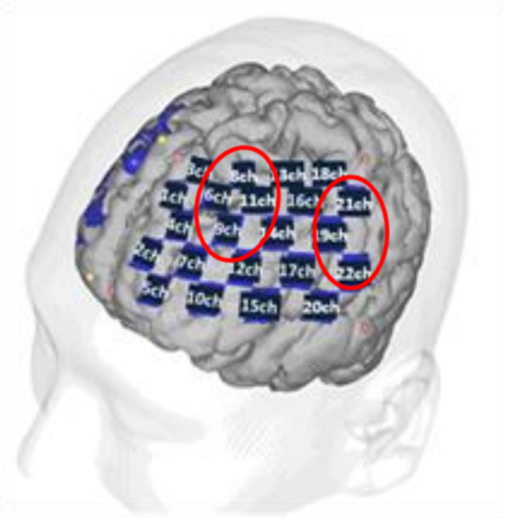


Fig.4. Brain activity of the statistical significant different.

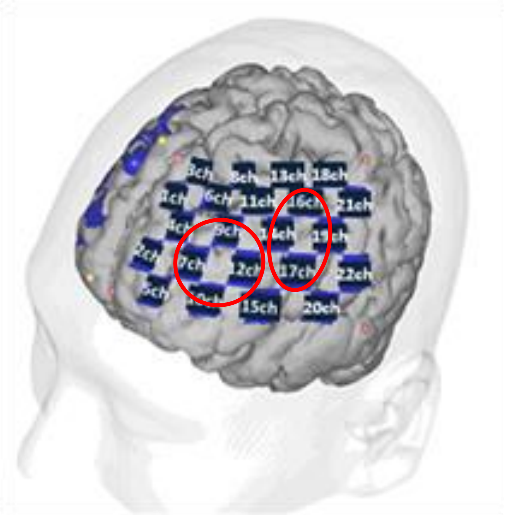


Fig.5. Significant differences when subjects turn left.

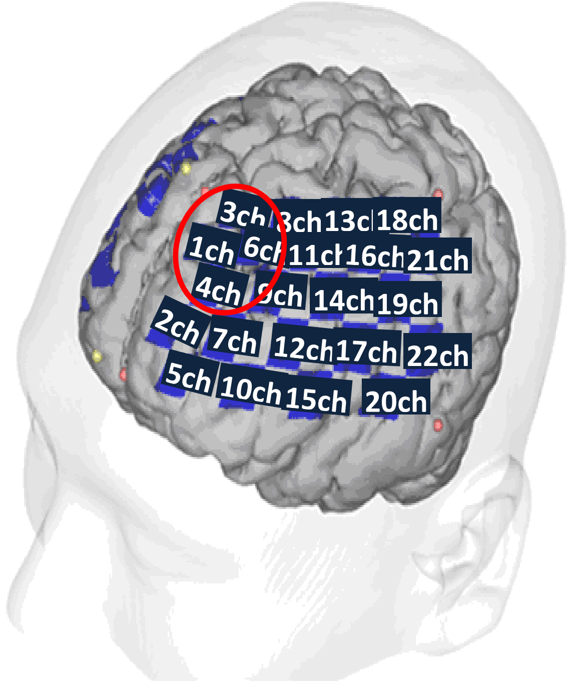


Fig.6. Significant differences when subjects turn right.

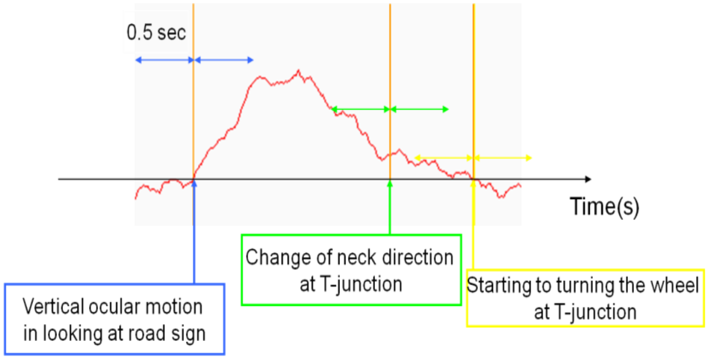


Fig.7. (a) Analysis method

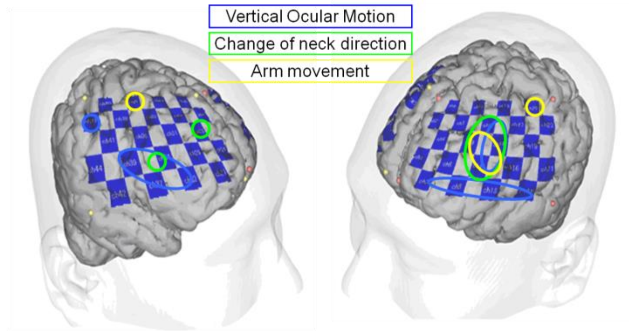


Fig.8. (b) Detail analysis when subjects turn left

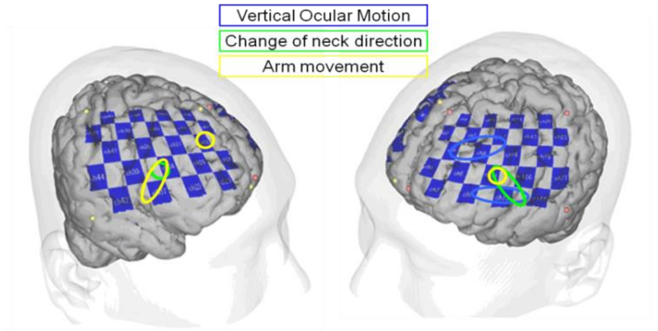


Fig.9. (c) Detail analysis when subjects turn right

IV. CONCLUSION

The hemoglobin density change of the human subjects' frontal lobe was partly observed in the experiments we designed, where three kinds of tasks were performed to analyze human brain activity from the view point of spatial perception.

The NIRS measures of hemoglobin variation in the channels suggested that human behavioral decision-making of different types could cause different brain activities as we saw in the tasks: 1) taking a given direction at the first T-junction, 2) taking a self-chosen direction on a road sign at the second T-junction and 3) turning the wheel or not. Some significant differences (paired t test) on NIRS's oxy-hemoglobin and less interrelated results between "pushing a button" and brain activity at the second T-junction are obtained.

Behavior related to the driving of the subject served to be the same condition as much as possible. It was not to use brain activity data when subjects performed different operation according to the situation of traffic. We think that differences in instruction and the difference between the turning left and right are reflected in the results of t-test.

In the analysis, t-test was performed on the condition that sample time length is 4 seconds or 1 second. In the general, it was said that hemoglobin density changes after a few seconds from being given stimulus. In the case of driving, subjects have to process various information and perform to operate car driving. If sample time length is long, it means that results of t-test include various information. By using the data of 1 second, we think that the target of t-test was limited only to the one brain activity.

In previous study, we performed one sample t-test above condition too, because we verify the results of the previous and this study.
Furthermore, experimental results indicated that with the subjects moving their hand in circle, regardless of right or left, 1) the same response was observed in the prefrontal cortex and premotor area, and 2) different patterns of brain activities generated by moving either hand clockwise or counterclockwise.

The regions observed were only those with the 5% and less significance level. Possible extensions could be applied to other regions with the 10% and less significance level for the future study. With a larger number of subjects, brain activity patterns need to be made clear. In addition, it is thought to take particular note of participation concerning working memory when car is driven.

Furthermore, it was found that there were significant differences around #44-45 area. It is well known that this region is corresponding to language area. So, it is thought that subjects look at road map to determine direction that they have to go according to word described in road sign.

From results of these experiments, there was significant difference around working memory. So, experiments focusing on relationship turning wheel and working memory will be performed. On the other hand, experiments as to actual driving were required a broad range of perception and information processing. Especially, subjects had to determine behaves depending on various information at T-junctions, that is, the color of the traffic light, presence or absence foot passengers and so on. And so, we plan to perform more static experiments. we attention to differences on the basis of turning direction and dominant hand. In addition, we will conduct the experiments in which subjects were narrowed down to left-handedness. Furthermore, researches into other human brain activities than spatial perception are to be necessary with accumulated data from fMRI (functional magnetic resonance imaging), EEG (Electroencephalogram), etc.

When compared virtual result to actual ones, there were significant differences around #46 area in both experiments, which were performed in virtual and actual condition, as a common result. It is thought that this result is due to activities of working memory because subjects must to recall memories of movements required for car driving and turning steering wheel. Conversely, there were significant differences around #10 in virtual experiments and around premotor area in actual driving, respectively. In the virtual case, it is thought to result from inhabitation of task without movement. In the actual case, subjects had to perceive space information in real time. So, it is considered that there were significant differences around premotor area because they always ready up to manipulate steering wheel.

In addition, we performed one sample t-test based on subjects' movement as detail analysis. Particularly, we paid the attention to the movement of the neck which confirmed road reputation, the lateral movement of the neck in the crossing, movement for steering wheel operation of the hand and performed one sample t-test. As a result, there is the significant difference on the difference brain region in each t-test analysis. Also, these results were difference from previous results that performed one sample t-test by the brain activity data when subjects were taught the direction or names of places. All the viewpoints are human movement about driving car in the T-junction when we analyzed it in detail. Thus, we think that decision making appears for the result of the expression point of the statistical significant difference not developing by movement when I decide a direction.

V. Future Work

As a future plan, we aim to apply these results to assistive human interface. As a matter of course, we plan to performed additional experiments including the verification of these results. And final purpose is to develop a new system for manipulating wheelchair and information presentation system to assist recognition of information including spatial one during car driving. Recently, some systems to assist the driver by using a lot of sensor have been developed in some car company. We think that these results are considered to develop some assist systems. For example, we think that the system which prevents mistakes of human decision and driving assist system to using BMI. Therefore, it is necessary to measure more the brain activity data and make the database.

References

[1] M.J. Kahana, R. Sekuler, J.B. Caplan, M. Kirschen, and J.R. Madsen: "Human theta oscillations exhibit task dependence during virtual maze navigation.", Nature, 1999, 399, pp. 781-784.

[2] N. Nishiyama and Y. Yamaguchi: "Human EEG theta in the spatial recognition task", Proceedings of 5th World Multiconf. On Systemics, Cybernetics and Informatics (SCI 2001), Proc. 7th Int. Conf. on Informatiofn Systems, Analysis and Synthesis (ISAS 2001), pp. 497-500 (2001).

[3] E.A. Maguire, N. Burgess, J.G. Donnett, R.S.J. Frackowiak, C.D. Frith, and J.O' Keefe: "Knowing Where and Getting There: A Human Navigation Network," vol. 280 Science 8/may/1998.

[4] Chin-Teng Lin, Shi-An Chen, Tien-Ting Chiu, Hong-Zhang Lin, and Li-Wei Ko: "Spatial and temporal EEG dynamics of dual-task driving performance." Journal of NeuroEngineering and Rehabilitation, vol. 8-11, 2011

[5] S. Shimizu, N. Hirai, F. Miwakeichi, and et al: "Fundamental Study for Relationship between Cognitive task and Brain Activity during Car Driving," Proc. the 13th International Conference on Human-Computer Interaction, (San Diego, CA, USA, 2009), Springer Berlin / Heidelberg, 434-440.

[6] N. Takahashi, S. Shimizu,Y. Hirata, H. Nara, F. Miwakeichi, N. Hirai, S. Kikuchi, E. Watanabe, and S. Kato: "Fundamental Study for a New Assistive System during Car Driving," Proc. International Conference on Robotics and Biomimetics, 2010, China.

[7] N. Takahashi, S. Shimizu, Y. Hirata, H. Nara, H. Inoue, N. Hirai, S. Kikuchi, E. Watanabe, and S. Kato,"Basic study of Analysis of Human Brain Activities during Car Driving," the 14th International Conferrence on Human-Computer Interaction, 2011, Orlando, Florida, USA.

[8] S. Shimizu, N. Takahashi, H. Nara, H. Inoue, and Y. Hirata ,"Fundamental Study for Human Brain Activity Based on the Spatial Cognitive Task," the 2011 Internatinal Conference on Brain Informatics-BI 2011, China.

[9] S. Shimizu, H. Nara, N. Takahashi, H. Inoue and, Y. Hirata,"Basic Study for Human Brain Activity Based on the Spatial Cognitive Task," The Third International Conference on Advanced Cognitive Techonologies and Applications, 2011, Italy.

[10] J. Cockburn: "Task interruption in prospective memory: "A frontal lobe function?." Cortex, vol. 31, 1995, pp. 87- 97.

[11] E. Watanabe, Y. Yamashita, Y. Ito and, H. Koizumi, "Non-invasive functional mapping with multi-channel near infra-red spectroscopic topography in humans," Heurosci Lett 1996, Feb 16, 205(1), 41-4.

A Smartphone Intervention for Cycle Commuting

Yun-Maw Cheng
Institute of Design Science and Department of Computer Science and Engineering
Tatung University
Taipei, Taiwan

Chao-Lung Lee
Department of Computer Science and Engineering
Tatung University
Taipei, Taiwan

***Abstract*—For those who new to cycling to and from work, how do you inspire yourself to keep up with it? Previous research has identified that having a partner is crucial to create and maintain a new habit. However, the rapidly changing work dynamics pose a challenge on this basis. Frequently failing to show up at the appointed or expected time can cause the motivation breakdown. In this paper, we introduce BikeTogether, a smartphone app that encourages and supports its users to cycle home with each other over the Internet. The app employs the metaphor of a bicycle flashlight to represent closeness, leading, and following between two sides. The cycling performance is also recorded so the users can track how they are doing over time. 10 participants were instructed and randomly paired to take a two-phases test ride on different routes. Results indicated that the app can help create the sense of being with each other while cycling and promote not only accompanied but competing ride. In addition, the outcome of the desirability towards the app implies a higher chance it will lead to a behavior change. This provides a new way that we can commit to remain encouraged.**

***Keywords*—copresence; behavioral nudge; persuasive computing; smartphone app**

I. Introduction

Fast paced lifestyle has compelled people to strike a balance between their work and health. Research has shown that exercising after work followed by getting a good night's rest is the essential to avoid physical and psychological exhaustion [1]. Cycling home after work can be a minimalistic approach that accomplishes the two tasks with only one effort [2]. Along with the benefit of the idea, people can simply make a positive change then repeat. However, this is easier said than done. The change of this kind is consequences of everyday decisions that assemble over time. Also, as the social comparison theory states: "Humans have a drive to assess how they are doing. In order to assess how they are doing they seek standards in which to compare themselves. When objective standards are not available, they look to their social environment and engage in comparison with available others" [3]. Social influences and supports play a critical role in encouraging and retaining the change [4]-[6].

Can the pervasiveness and technological capabilities of smartphones be designed to enable the change? As we have seen, they are not only for communication but also for measurement of variations in daily activities (i.e. status update, check-in, etc.). The use of smartphones together with various sensors is attracting attention of cycling communities. The information they provide is mainly about fitness and performance. The fitness information is about a cyclist's physical status, such as heart rate, pulse oximeter, skin temperature, galvanic skin response, or calories burned. The performance information is regarding a cyclist's cycling conditions, such as speed, time, and distance traveled. The information is then forged into the process of goal-setting, self-monitoring, rewards and sharing features in order to better the cycling experience.

With all these, there are possibilities that can empower people to engage in collecting and sharing relevant real-time information to foster collective action based on their measurements [7][8]. To leverage the materialities of smartphones to embody the effect of social comparison on motivation for cycle commuting, there are questions to be answered [9][10]. What data are collected? How are the data turned into appropriate information and presentation, which is focused and meaningful enough for cycle commuters to understand and help them to make choices to hang on to the change?

These commuters get on their bicycles and let their legs carry them home. The exertions are transformed into acceleration and distance ridden. Mounting a smartphone on the handlebar to track these locomotion details and to share them between the two commuters in real-time is as handy as it is becoming. The sharing is considered a key in creating a sense of togetherness [11]. Also, let the information be contextual can support each of them to focus on their cycling without the necessity to reason it cognitively [12] [13]. The challenge herein is the selection of a suitable communication medium with its visibility and glanceability to carry the information in situ. Cycling home after work reminds us the importance of lighting. In addition to safety concerns, feeling isolated and alone is what augments the decreasing of the motivation. Lighting can create atmosphere and mood that makes us feel guided and accompanied [14][15]. Moreover, as Fogg suggests, flashing lights is one of the form factors that can motivate people and prompt immediate action [9]. Utilizing the screen of a smartphone as the lighting source to convey the exertion information in an ambient, unobtrusive, and intuitive manner could have the potential toward the goal of influencing the commuters to be more active.

Most people are already rely on the Internet as a source and a tool to broaden their circle of social interaction to include possibly everyone who share the same interests, beliefs, and goals. The mobility of smartphones can offer support for sensory extension to enable remote individuals to continue interacting with one another while moving around. Using dynamic patterns of light, such as color and intensity, on the screen of a smartphone as sensory cues to define the relationship between the cycle-commuters' perception and their cycling activity provides a tangible approach in creating a remote sense of togetherness. In response, we developed BikeTogether, an app to explore the possibility of using smartphones mounted on the handlebars as an ambient display to convey information that embodies social comparison to the commuters engaged in their cycling. To evaluate this experience, a series of studies were conducted to realize how those new to cycle-commute interpreted and responded to the design. In brief, this research makes the following contributions:

- It identifies that smartphone apps can be designed to create a sense of togetherness between two cycle-commuters while cycling. This leads to positive behavioral effects.
- It identifies what and which visualization and where to present information using a smartphone while cycling to embody remote presence.
- It confirms the feasibility of this approach through experiments on our prototype implementation via an in-the-wild study. The experimental results show that the app exhibits a good degree of persuasion.

II. Design Goals and Choices

BikeTogether incorporates social comparison theory with the materiality of smartphones and ambient display. This section describes the design goals we want to achieve with our design choices. In order to know more about the motivation behind cycle-commuting, a brief interview of 5 colleagues in our department who cycle-commute at least on a part-time basis (1 female, 4 males, M = 54.6 years, SD = 3.92) was performed. The results showed that getting home quicker and feeling of healthier in terms of psychological, physical, and environmental outcomes hold prime importance in their choice of the commuting option. All of them expressed the feeling of wanting others to share in their enthusiasm. Also, a street interview of 15 random people (6 female, 9 males, M = 36.86 years, SD = 13.7) waiting in line for a bicycle at a YouBike station, which is the public bicycle rental scheme in Taipei, during after work rush hour revealed reasons for cycling home other than what the colleagues expressed. Their responses included: it is cheaper because the YouBike provides free for the first 30 minutes promotion; I want to fit in the popular crowd with the idea of leaving the environment in a better condition. These reasons suggested that it has potential to encourage people to cycle-commute and keep up the change through the conversation with the like-minded others in situ.

There may be people with the same interests around can help keep us motivated. But we do not bump into each other as we expected to be. So go beyond our naked sense perceptions to reach each other in order to get motivated and inspired to do something that matters together is the challenge that the design of BikeTogether attempted to address. Those cycle-commuters are related by what they share in common but unknown to each other. The app is aimed at creating a form of conversation that enables its users to freely express to each other in the comprehension that they might never meet again [16][17]. This comprehension, however, intends to intrigue the thought of who and where the users really are to maintain the surprising freshness. Therefore, the level of anonymity the app adapts is that the users' locations are veiled while their identities are disclosed.

As mentioned, the app is designed to create a sense of togetherness using light as a communication medium regardless of the users' physical distance. The sense of being together with each other is a subjective experience. To fabricate this illusion, a synchronous communication between the two sides is needed. Also, finding intuitive and accessible manipulations in situ to reconfigure those preconceptions of smartphone use through the ambient engagement of the viewer and the use of visual sensory mapping strategies should be adopted. The strategies would then position the users as active producers of meanings to each other. This draws attention to the ways in which smartphones can bridge the two sides of users who want to do something together but cannot find each other in an implicit and unobtrusive manner. Along with this idea, there are also the questions of what the core messages in the conversation is and what the concept they mutually agree.

People have a radical desire to assess their ability [3]. It is, as a matter of course, the levels of exertion measured by acceleration and cycling distance are chosen as the core messages. Dubberly and Pangaro argued, "Conversation is a progression of exchanges among participants. Each participant is a 'learning system,' that is, a system that changes internally as a consequence of experience" [18]. The idea of BikeTogether is to provide its users with an informative and motivating representation of the core messages using light based on a self-reflective approach.

The light intensity is to be proportional to the acceleration during a cycling activity. In order to inform the users about how they are doing during their cycling, cumulative values of the distance ridden so far can be summed up, synchronously exchanged between the two sides of the users, and displayed in an ambient manner as the current ongoing overall result on the screen of the smartphone as colors. A green, for instance, could signify that the user has obtained a leading position compared to the other user, while a red would indicate that the user is on following position because the cycling distance is less than the user's on the other side. An orange indicates the closeness of two sides of the users.

The key thing about cycle-commuting is not simply to set a goal and to turn into a self-starter instantly. It has to be accomplished in a way that people feel they can now manage things they want to do better than before. BikeTogether aims to make it possible for the users to have a feeling of "I have a chance to meet a new partner every time when I cycle home". The feeling of togetherness, therefore, purposely encourages the users towards a desired behavior. Another deliberation for the users is to provide cumulative information about the total cycling distance since the app is taken on. In some way, it encourages the users to compete against themselves. This not only helps them reflect on cycle-commuting which is something they build up to but also the effect of feeling of togetherness on the rate of progress they perceive. The results may lead to develop from a personal need for achievement to an interactional and varying achievement goals. As to the design, the metaphor of a car mileage meter is used to describe the increase that occurs at each step of the progress. Over time, after glancing and contemplating the numeric difference, the sensation that we have a hope of accomplishment with like-minded others may grow stronger [19][20].

III. Interactions of the BikeTogether

The goal of BikeTogether is to motivate people who just start cycle-commuting to do it more. In order to explore the design opportunities from the design choices discussed previously and to have a better understanding on how it carries out, the interactions are shown below:

1) The users simply grab their bicycles, mount their smartphones on the handlebars, and turn on the app. They then come to the selection screen, which is a flashlight with an on/off button as the only one major user interface element. The button is a circular menu with four features, “Double”, “Solo”, “History”, and “Setting”. The users can circle along the selection and press the on/off icon to confirm the selection.

2) When the app runs up for the first time, the user enters information such as identity and password, to create an account. This is all done by going to “Setting” in the circular menu.

3) When “Double” is selected, the outer edge of the button starts flashing in white and this indicates that the system is attempting to pair two sides of the users whoever are online. If it becomes solid white, this indicates that the connection is successfully established and the system starts to facilitate the exchange of distance data between the users since they are paired. Otherwise, the process is again repeated.

4) Also, the users can compete with their previous best times and efforts by selecting “Solo”. The connection with your virtual self is made instantly once the button is pressed.

5) The flashlight illuminates in different colors. Red, green, and orange lights denote following, leading, and closeness between the paired users in terms of the distance the users have ridden. The intensity is directly proportional to the level of acceleration and the flickering means the current positional state is about to change.

6) When the on/off button is pressed second time by either side of the users, the pair is ended immediately and a review of the current results is then brought up. The user can check distance, average speed, time (duration and date), and paired partner’s ID of the cycling. A route map is shown on the screen as well. The map displays the leading, closeness, and following distance, represented by green, orange, and red respectively.

7) The users can browse their cycling records by selecting “History” on the selection screen at any time. For a more elaborate look at any individual cycling activity, click the bar representing that cycling to review the route map and other details.

In the next section, an evaluation of the interaction and interface with users in the real scene is described. The hope is to discover missing elements or parts that can be smoother in encouraging people go for and keep up cycle-commuting.

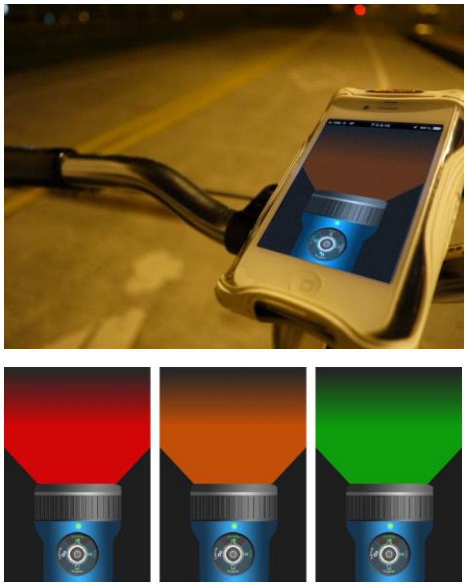
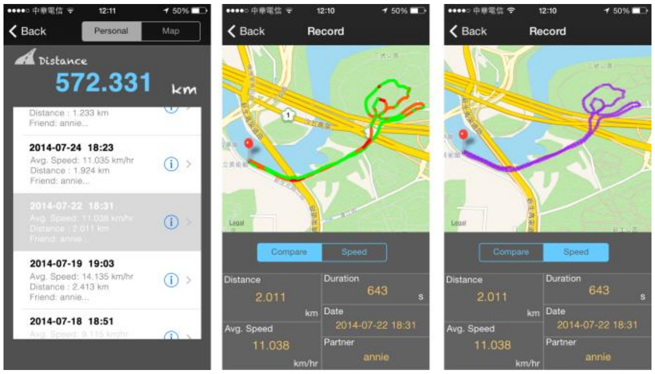


Fig. 1. BikeTogether in the field (top); Color changing of flashlight denotes the relative position, red: following, orange: closeness, green: leading (bottom).



Distance : 1.233 km
Friend: annie...

2014-07-24 18:23
Avg. Speed: 11.035 km/hr
Distance : 1.924 km
Friend: annie...

2014-07-22 18:31
Avg. Speed: 11.038 km/hr
Distance : 2.011 km
Friend: annie...

2014-07-19 19:03
Avg. Speed: 14.135 km/hr
Distance : 2.413 km
Friend: annie...

2014-07-18 18:51
Avg. Speed: 9.115 km/hr

Distance 2.011 km
Duration 643 s
Date 2014-07-22 18:31
Avg. Speed 11.038 km/hr
Partner annie

Distance 2.011 km
Duration 643 s
Date 2014-07-22 18:31
Avg. Speed 11.038 km/hr
Partner annie

Fig. 2. Cycling history and performance.

IV. Methodology

In order to understand if BikeTogether can effectively convey information to people engaged in a cycling activity via the screen light from a smartphone mounted on the handlebar, and whether this can influence their attitude towards the transit of this kind, a two-phase experiment was designed and carried out with real users in the Dajia Riverside bicycle path, which is one of the busiest paths for cycle-commuting in Taipei City. The participants consist of 10 undergraduate and postgraduate students at Tatung University (1 females, 9 males, M = 23.1 years, SD = 1.14). The majority of the participants (90%) use cycling as a recreation activity, and half of them prefer cycling with a companion. On the day prior to the scheduled experiment, participants received a notification via Facebook Messenger to remind and confirm the appointment. Also, they were all notified to wear comfortable clothing.

Two bicycles both with a smartphone mounted on the handlebars and BikeTogether preinstalled were prepared for the evaluation. The participants were then gathered, randomly split into 5 groups of two, and told to take a 10 minutes open cycle, which means they can ride on their own pace, on two different routes, one group at a time. As to how their cycling behavior is modeled, the mode of cycling is considered to be accelerating, decelerating, and constant speed in this evaluation. The cycling behavior of each participant therefore can be represented as an acceleration-time sequence within a certain time frame that meets the reality of cycling circumstances [21][22]. In this case, a time frame of 5 seconds was chosen.

These data were collected and calculated by the app ran in stealth mode without any GUI so the participants did not know they were being traced. After a 30 minutes rest, the participants were gathered once again and a brief instruction regarding the information coding of the app was given. They were then took another 10 minutes cycling with the app ran in normal mode, one group at a time. The data were again collected. Figure 3 shows the rate of acceleration and deceleration in both study phases. The specified threshold for count is ±2m/sec^2.

A Wilcoxon signed-rank test showed a statistically significant difference before and after participants were aware of BikeTogether ($W=10$, $P \leq 0.05$). Also, to better understand how the app effected on the participants, their relative position to each other in terms of the real-time cycling distance in the second phase was summarized and an interview was performed.

V. Results

All the participants reported the increase in their activity level and considered that it was mainly due to the more aware of their personal activity as well as the comparison with remote partner's provided by the app. Group 4 said that it started like a subtle competition and later it became more like a group ride. Group 1 and 2 expressed that, at first, they attempted to attack off the front and later figure out the different physical ability in between, then slow down to their moderate pace. A common group cycling formation was seen in Group 3 and 5. They commented that once in leading position they would lower their speed in order for the other peer to catch up.

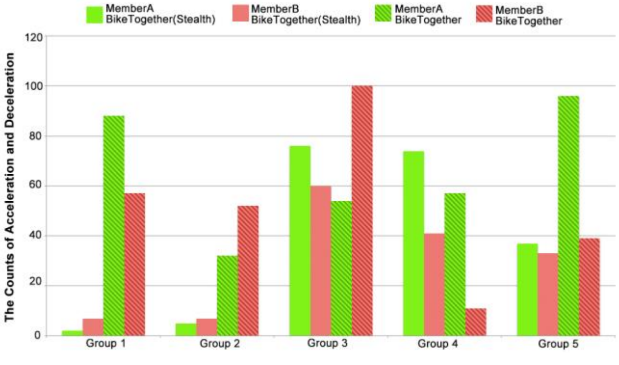


Fig. 3. The counts of acceleration and deceleration in each phase

In order to discover more about whether BikeTogether can engender positive effect that the users would use it frequently and longer, Production Reaction Cards was used to capture the participants' feelings and study their emotional involvement with BikeTogether [23]. These cards are composed of 60% of positive words and 40% of negative and neutral words. The participants were asked to select as many words as they wanted. The most selected word was fun (72%), followed by entertaining (70%), exciting (64%) and motivating (51%). All the participants also uttered that the total cycling distance directly provides them the feeling of achievement. This makes us have confidence in that BikeTogether can help further enhance the cycling experience and has potential to nudge people who just start cycle-commute to keep up the change.

VI. Conclusion and Future Work

Hard to find mutually convenient time and place with cycling partners is the common cause that hinders turning it into habit. The findings from this research indicate that smartphone apps can be designed to create a sense of togetherness between two cycle-commuters while cycling using the screen light. This is a form of instant messaging, a bicycle-to-bicycle, body-to-body conversation. It is even more like a form of networked, peer-to-peer augmented game. The one you chat and compete with, right at the moment, numbs your orientation in time. This leads to positive behavioral effects. The next stage will involve the usability tests and user experience evaluation of the prototype to see whether the increased feelings of presence increase motivation in the long term so that the health and fitness goals will be achieved?

Acknowledgment

This research is partially funded by the Ministry of Science and Technology grand 101-2221-E-036-010. Thanks to all our participants for taking part in this study and sharing your inspirations.

References

[1] I. J. Nägel and S. Sonnentag, "Exercise and Sleep Predict Personal Resources in Employees' Daily Lives," Applied Psychology: Health and Well-Being, vol. 5, no. 3, pp. 348-368, October 2013.

[2] S. Handy, Y. Xing, and T. Buehler, "Factors associated with bicycle ownership and use: a study of six small U.S. cities," Transportation, vol 37, no. 6, pp. 967-985, Nov 2010.

[3] R. P. Larrick, K. A. Burson and J. B. Soll, "Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)," Organizational Behavior and Human Decision Processes, vol. 102, pp. 76-94, December. 2006.

[4] A. S. Gabriel, J. M. Diefendorff, and R. J. Erickson, "The relations of daily task accomplishment satisfaction with changes in affect: A multilevel study in nurses," Journal of Applied Psychology, vol. 96, pp. 1095-1104, September 2011.

[5] M. Kanning and W. Schlicht, "Be active and become happy: An ecological momentary assessment of physical activity and mood," Journal of Sport and Exercise Psychology, vol. 32, pp. 253-261, April 2010.

[6] E. Massung and C. Preist, "Normification: using crowdsourced technology to affect third-party change," In CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13), pp. 1149-1154, New York, ACM Press, 2013.

[7] S. Heggen, "Participatory sensing: repurposing a scientific tool for STEM education," interactions, vol. 20, no. 1, pp. 18-21, January 2013.

[8] J. van der Linden, Y. Rogers, and V. Waights, "A blended design approach for pervasive healthcare: Bringing together users, experts and technology," Health Informatics Journal, vol. 18, no. 3, pp. 212-218, September 2012.

[9] B. J. Fogg, D. Eckles, Mobile Persuasion: 20 Perspectives on the Future of Behavior Change; 1st ed.: Stanford Captology Media, 2007, pp. 77-84.

[10] T. R. Chang, E. Kaasinen, and K. Kaipainen, “What influences users' decisions to take apps into use?: a framework for evaluating persuasive and engaging design in mobile Apps for well-being,” In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM '12), 2012, pp. 2:1-2:10.

[11] S. Zhao, “Toward a Taxonomy of Copresence,” Presence: Teleoperators and Virtual Environments, vol. 12, no. 5, pp. 445-455, October 2003.

[12] V. Occhialini, H. van Essen, and B. Eggen, “Design and evaluation of an ambient display to support time management during meetings,” In Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction (INTERACT'11), ed. P. Campos, N. Nunes, N. Graham, J. Jorge, P. Palanque, pp. 263-280, vol 2, Berlin: Springer-Verlag, 2011.

[13] A. Göker, S. Watt, H. I. Myrhaug, N. Whitehead, M. Yakici, R. Bierig, S. K. Nuti, and H. Cumming, “An ambient, personalised, and context-sensitive information system for mobile users,” In Proceedings of the 2nd European Union symposium on Ambient intelligence (EUSAI '04), pp. 19-24, New York, ACM Press, 2014.

[14] Y. Rogers, W. R. Hazlewood, P. Marshall, N. Dalton, and S. Hertrich, “Ambient influence: can twinkly lights lure and abstract representations trigger behavioral change?,” In Proceedings of the 12th ACM international conference on Ubiquitous computing (Ubicomp '10), pp. 261-270, New York, ACM Press, 2010.

[15] C. Harrison, J. Horstman, G. Hsieh, and S. Hudson, “Unlocking the expressivity of point lights,” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), pp. 1683-1692, New York, ACM Press, 2012.

[16] H. C. Stuart, L. Dabbish, S. Kiesler, P. Kinnaird, and R. Kang, “Social transparency in networked information exchange: a theoretical framework,” In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 451-460, New York, ACM Press, 2012.

[17] R. Kang, S. Brown, and S. Kiesler, “Why do people seek anonymity on the internet?: informing policy and design,” In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), pp. 2657-2666, New York, ACM Press, 2013.

[18] H. Dubberly and P. Pangaro, “On Modeling: What is conversation? How can we design for effective conversation?,” interactions, vol. 16, no. 4, pp. 22-28, July 2009.

[19] J. McGonigal, Reality Is Broken: Why Games Make Us Better and How They Can Change the World, Penguin Books, 2011, part I.

[20] P. Lunenfeld, A. Burdick, J. Drucker, T. Presner and J. Schnapp, Digital_Humanities; the MIT Press, 2012, ch. 3-4.

[21] J. Dill and J. Gliebe, “Understanding and Measuring Bicycling Behavior: a Focus on Travel Time and Route Choice,” Report, Oregon Transportation Research and Education Consortium (OTREC), 2008.

[22] K. J. Krizek, S. L. Handy, and A. Forsyth, “Explaining changes in walking and bicycling behavior: challenges for transportation research,” Environment and Planning. B, Planning and Design, vol. 36, no. 4, pp. 725-740, January 2009.

[23] J. Benedek and T. Miner, “Product Reaction Cards,” Microsoft, July 29, 2002.

Broadband Access for All: Strategies and Tactis of Wireless Traffic Sharing

Jinadu Olayinka T.
Department of Computer Science,
Rufus Giwa Polytechnic, Owo,
Ondo State, Nigeria

Owa Victor K..
Department of Computer Science,
Rufus Giwa Polytechnic, Owo,
Ondo State, Nigeria

***Abstract*—Network engineers have designed an array of protocols that enabled shared access in various wired and wireless contexts at different layers of the protocol stack [1]. One approach to managing unlicensed spectrum is to rely on a technical protocol to allocate and manage shared access Lehr [2]. This paper addressed the benefits of unlicensed wireless traffic within licensed traffic (anticipated as cognitive router-based networking). A focus on shared access to non-exclusive use of the spectrum with an holistic view of technical and institutional features is suggested for effective management of 'spectrum commons'. Using the adapted cognitive radio architectural model and its associated multi-hop ad-hoc networking strategies to implement the 'spectrum-common', mobility is enhanced with each node acting as a router and packet forwarder. We formulate management frameworks that can integrate well with liquid protocols for mobile nodes. Also, these frameworks incorporate new strategies of intelligently adapting the nodes to dynamically participate in setting bandwidth capacity stochastically. The projected use of dynamic bandwidth shaping algorithm for the cognitive radio-based network (CRN) when implemented will make broadband access more economical to users and the spectrum used effectively.**

Keywords—ad-hoc; spectrum commons; etiquettes; software defined radio (SDR); MIMO

I. Introduction

The significant progress in wireless technology and the growth of wireless services has provided the principal impetus for reforming spectrum management and hence the transition toward increased reliance on market forces. While many wireless technologies contribute to both the viability and desirability for managing spectrum via unlicensed (smart wireless system technology, including software or cognitive radios, smart antennas and multiple input multiple output MIMO system) platforms, the benefits of unlicensed wireless are best anticipated in the context of ad-hoc networks [3].

The ad-hoc networks are mobile, dynamic wireless networks that require no fixed infrastructures [4]. As the continuous end-to-end connectivity between its mobile nodes is not guaranteed, [5] pointed out that the ability to self-form and self-mange remains a major challenge. Due to this partial and intermittently connected wireless frameworks, the mobile ad-hoc network (MANET) hosts induce link disruptions, which may result to degradable service disruptions except assisted by derived technologies including intelligent etiquettes and strategized mobility management.

Today, spectrum licenses to provide mobile services offer an entry barrier that gives incumbent licensee a strategic advantage. However, with robust competition and the threat of increased allocation for competing wireless technologies on one hand and the prospect of having to pay for additional spectrum to support new (3G wireless broadband) services on the other, the mobile operators are more inclined to share spectrum [1] and [2].

A. Motivation

As policy-makers are committing to a dual regime of flexible licensed and unlicensed spectrum to provide for the evolution from the centralized approaches to more decentralized management regimes, the elements of a protocol for managing the spectrum commons must be defined. These new protocols are required both at the level of running codes (as protocols and standards) and at the level of institutional frameworks.

Also, as wireless traffic become more like Internet traffic with heterogeneous, bursty or fat-tailed, long-hold time for connectivity but variable link status due to ad-hoc networking, there is need to deploy now strategies to manage wireless resources [6]. The proposed rules was examined for expected performance and support for ad-hoc communications with reduced overheads but increased quality of service (QoS) in [1].

B. Objectives

The objectives of this research are to:

- define suitable framework and CR-based infrastructure for a spectrum commons;
- incorporate learning strategies to make the defined protocols liquid and
- suggest approaches of incorporating defined etiquettes into existing management protocols to achieve sharing goals.

II. Review of Related Works

A. Regulatory Models

Reference [2] identified three models of spectrum management to include commands and control (C&C), property rights (as licensed) and open access (as unlicensed) users.

As discussed, C&C is a scheme whereby the government acts as the regulatory agency such as the Federal Communication Commission (FCC) in the US or Ofcom in the UK. Here, the government controls the choice of technology, spectrum uses and users. According to [7], this system is vulnerable to influence costs. As the government regulators lack the expertise to make informed decisions, the regulation is often slow and expensive, and therefore it is criticized as non-market-based approach [8].

In contrast, the licensed (property rights or exclusive-flexible use) and unlicensed (open access or 'commons') models are approaches stylized as market-based because the decision making power is decentralized to the market. In these schemes, the service providers, equipment makers and end-users interact and compete in the market place to determine spectrum usage.

Reference [7] further explained that even as the licensed scheme confers a property right on the licensee to use the spectrum exclusively, there are rules, which limits its tradability and licenses are subject to term limits. In the same vein, the assumption of relative spectrum abundance is provisioned by the unlicensed scheme. Reference [9] also corroborated the unlicensed model as an open access scheme operated as a 'commons' approach, where the right to access or use the spectrum is shared among users. though, under the licensed approach, an exclusive use license assigned may be traded in secondary markets, and licensees only have flexibility in the choice of technology and services offered. In addition, licensees are just allowed to trade the usage rights conferred by the license.

As the commons approach provide the right to access the spectrum in a shared manner (among the users) subject to protocols, the decision-making authority is decentralized to those who share access to the commons, and as the protocol embodies the mechanism for managing the spectrum, the decision –making is governed by the protocol put in place. Moreover, much flexibility is offered in commons even as the choice of the protocol may be made by the government or by the market via industry standardization unlike in license regime, where decision-making resides only with the central planner (government).

However, the 'commons' approach does not suggest that the spectrum will be free but it is open access to only those who conform to the unlicensed protocol. Furthermore, the unlicensed does not mean unregulated as costs incurred will be borne by users, either directly through access payments or indirectly through taxes, protocol implementation cost or congestion-related quality of service effects. These costs include costs of setting up and operating the management procedures such as processing costs to implement sharing protocol, its enforcement and control congestion. This is also borne in license.

Several additional distinctions between the licensed and commons are noted in [2]. They are both 'shared' in the sense that multiple devices and end-users simultaneously access and use the spectrum. For example, mobile operators share spectrum over multiple users, and competition among operators offers competition across technologies and markets. Also, they are both market- based and as these models offer dynamic spectrum access and movement by end-users via roaming and switching among operators, the mobile customers are secondary licenses who get to use the spectrum on the basis of rules established by the licensed operators

B. *Communication Standards*

The standard for modern telecommunication networks is to offer 99.999% availability. References [10], [11], [12] and [13] all discussed the role of unlicensed (commons) regimes as sure step towards providing solution to spectrum scarcity and a promoter of innovations in telecom services.

The rules for managing a spectrum commons as stipulated in [12] and [14] showed that centralized resource allocation mechanisms (ATM, token ring) provide more assurance of bounded access delays while distributed protocols (TCP, Ethernet) provide similar delays when networks are lightly loaded. Centralized approaches are less robust in dynamic state of ad-hoc networks, which characterise future wireless environments.

Similarly, VolP perfectly co-exist with FTP, email and other data traffic when the network is not congested. With TCP and IP segmentation of packets in transport, IP hop-by-hop and TCP (end-to-end) provides the special controls of allowing packets in variables length. As remarked in [2], much of the licensed spectrum (ISM band) used by Wi-Fi, Wireless LAN or Bluetooth is managed in a decentralized way analogous to the Internet and the applications are adaptive making resources isolation less strictly managed.

For these and many other standards to be effectively upheld to provide broadband access for all and BGP providing inter domain routing support, a more decentralized approach may be the only feasible way to manage resources. This also includes decoupling of spectrum frequencies from infrastructure investment and applications.

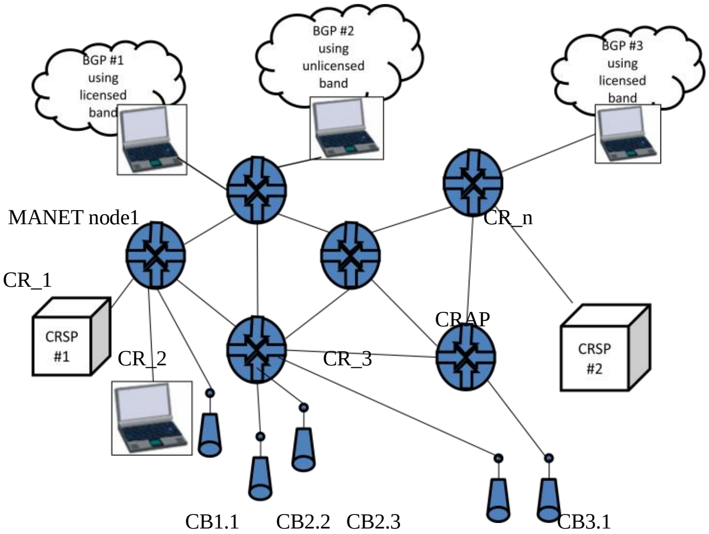
III. Design Framework for 'Spectrum-Commons'

The design of an appropriate framework for managing unlicensed spectrum is conceived to be minimally constraining but very consistent with orderly management of shared access spectrum.

Development of framework or rules structured for operating unlicensed devices to co-exist with licensed devices as primary users in dedicated unlicensed spectrum is crucial to the sharing.

A. *Spectrum Sharing Platform*

The environment of mixed regimes as (fig.1) provides for bulk of spectrum allocated via licensed and market-based unlicensed use. With cognitive radio network architectures and the dynamism exhibited by ad-hoc networking, the framework model is evolving, promoting innovations, and minimizing regulatory distortions. The design support marginal adjustments between licensed and unlicensed users, and within unlicensed supporting all changing protocols as need arises [7].


BGP #1 using licensed band
BGP #2 using unlicensed band
BGP #3 using licensed band
MANET node1
CR_n
CR_1
CRSP #1
CR_2
CR_3
CRAP
CRSP #2
CB1.1
CB2.2
CB2.3
CB3.1

Key:
CRAP – Cognitive Radio Access Point
CB – Cognitive Base
BGP – Border Gateway Protocol
CRSP – Cognitive Radio Service Provision
MANET – Mobile Ad-hoc NETwork node

Fig. 1. Prototype Mesh model for CRSP (adapted from CRSN architectures)

The proto-typical design includes licenses and unlicensed bands running BGP and the radio systems made smarter. This architecture enables dynamic spectrum sharing and the framework favours distributed/decentralised management characterised with maximal ‘common’ benefits. Reference [9] posited set of etiquettes as rules and mechanism to instantiate a common regime. It includes ‘protocol’ of running code for a software radio and technical standards for guiding the protocol design for a closed common.

Fig. 1 depicted a ‘closed spectrum-common’ platform for licensed and qualified operators (spectrum users) to implement the management regime for spectrum usage efficiency. A collective ownership of 3G spectrum and its management regime is prototyped as a ‘closed spectrum common) in this paper.

B. Design Rules

In agreement with [15] and [16], an infrastructural framework proposed to support the traffic sharing under secured Internet routing defined by BGP is characterised with:

- technology and associated capabilities to counter communication problems such as interception, interference, eavesdropping, spoofing, jamming, data falsification etc);
- frequency agility, expanded capacity for sharing, no transmit only device spreading spectrum capability and , transition to broadband platform;
- network provisioning for bursty traffic, multimedia services and other profiles;
- heterogeneous network technology provisions 3G, Wi-Fi, Infrared, satellites roaming and seamless mobility and
- spectrum reform policies, transits to expand flexible licensing and unlicensed spectrum management regimes instituted and sustained by defined etiquettes

IV. Implementation of a Liquid Protocol

Wireless traffic control schemes for broadband services includes constant bit-rate (CBR), variable bit-rate (VBR) unspecified bit rate (UBR), guaranteed frame-rate traffic flow (GFR) and available bit-rate (ABR) service categories [17].

For liquidity, the available bit-rate (ABR) scheme is envisioned to work in the spectrum commons. ABR scheme is capable of dynamically adjusting to the varying bandwidth capacity. The bandwidth made available to an ABR connection on any link varies between minimum cell rate (MCR) and the peak cell rate (PCR).

The learning automation is a sextuple defined in [17] as

<A,B,P,T,G,E>

Where

$A = \{a_1,a_2,...,a_r\}$ is the set of r actions offered by environment

$B = (0,1)$ the input set of possible environmental responses.

Q = set of possible internal states of the automation

P = probability distribution over set of action $P(t) = \{P_1(t), P_2(t), ... P_r(t)\}$ $Pi(t)$ is the probability of selecting action $a \epsilon A$ at time instant t_i.

$G = Q = A$ is the output function. G is deterministic one-to-one function.

E = estimator containing environmental characteristics

Using the learning automation, an estimator stochastically computes the output function obtainable in a bandwidth usage environment for a set of possible environment responses on the radio. With the CRN models $R_{1,}...,R_n$, the estimator is updated. Consequently, it adapts to environmental changes such as ABR bandwidth and is implemented by the learning algorithm T presented in fig. 2. It is used for obtaining the estimator vector

$$U(t)=U_1(t),U_2(t),...,Ur(t) \qquad (1)$$

Where

$U(t)$ is the estimator vector at any time instant t, and

$$E(t)=D'(t),M(t)U(t) \qquad (2)$$

Where

$D'(t)$ is the deterministic estimator vector at any time t; $D(t) = \{d'{,}(t), d'_2(t),...,d'r(t)\}$ and

$M(t) = \{M_1(t), M_2(t), Mr(t)\}$ is oldness vector; $Mi(t)=t-\max\{j : j<t; a(j)=ai\}$

Combining equations (1) and (2), algorithmic description of T is given in fig. 2.

```
Initialization: all  $P_i = 1/r$
Step 1:  Select an action $a(t) = a_k$
Step 2: Get feedback $b(t) \epsilon (0,1)$ from environment
Step 3: Computer new deterministic estimate $d_{ik}(t)$
Step 4: Update oldness vector $M_i(t) = m(t-1) + 1$
Step 5: For every $a_i$ $(i = 1,2,...,r)$, compute new estimate $U_i (t)$
Step 6: Select optimal action $a_m$ with highest estimate $U_m = \max \{U_i(t)\}$
Step 7: Update the probability vector
```

Fig. 2. Learning Algorithm for Automation

A. Discussion on CR–based model

The physical architecture of cognitive radio (CR) in ad-hoc setups make it feasible for receiving wideband signal. As software defined radio (SDR), with its radio frequency (RF) frontend, it is equipped with the capability to detect any weak signal in large dynamic range. This communication model can tune to any frequency band to receive any modulation.

As the estimator will be updated by the ABR connection source parameters – bandwidth resources are reserved for CBR and VBR connection that will be set up and the bandwidth becomes free again when CBR and VBR connections are released. This non-reserved bandwidth made available to ABR connections make all traffic sharable.

B. Modalities for defined etiquettes

Using BGP routing protocol, the cognitive-based service network and its special feature integrates well with other routing protocols. BGP also enables routing across all Internet service and other network providers. Combining with other technologies (WLAN, spread signal, infrared, WiMAX etc), temporarily unused band is used by any of the opportunistic radio, based on defined etiquettes to improve overall spectrum utilization [18].

As the estimator is updated by ABR connection source parameters – bandwidth resource is reserved for CBR and VBR connections set up and it becomes free again when CBR and VBR connections are released. This non-reserved bandwidth made available to ABR connections make virtually all traffic sharable.

To evaluate the ‘commons’ management regime, the following application specifications supported in unlicensed spectrum, under well defined protocols:

- Wi-Fi model of unlicensed devices – promotes innovation in wireless devices and IT business;
- mobile operators sharing of 3G spectrum minimizes transaction costs for accessing spectrum individually;
- realization of community mesh networks – provides mechansms for managing congestions, emphasizing co-ordination in co-existence.
- reliance on industry standardization process – fosters spectrum–specific etiquettes of management since the ‘commons’ regime also require specialized mechanisms.

V. Conclusion

The capability of cognitive radio (CR) within the wireless traffic provides many of current wireless systems with adaptability to existing spectrum allocation and overall spectrum utilization. CR supports common channels signalling; enabled with consistent security and privacy, envisioned in secured BGP [16].

Also, the commons spectrum will be more attractive to applications, which are adaptive and reasonably tolerant to congestion [19]. The system therefore, having mechanism for allocating resources among users/uses is equipped with established procedures to verify protocol is in conformance with agreed etiquettes.

With licensed wireless environment there are increasing demand and use of heterogeneous devices, uses leading to relatively insufficient spectrum. Spectral usage will be more efficient and spectral scarcity alleviated for broadcast and communication networks if suggested model is adopted. Users will benefit more significantly. Strategies to enhance wireless mobility management for qualitative seamless roaming and service continuity are suggested for future research.

Acknowledgment

We are grateful to the management of Rufus Giwa Polytechnic, Owo, Nigeria and Tertiary Educational Task Fund (TETFUND) Nigeria for the academic staff training and development fund made available to publish this research paper.

References

[1] S. M. Benjamin, “Spectrum Abundance and the choice between Private and Public Control”, New York University Law Review, vol 78, 2007.

[2] W. Lehr and J. Crowcroft ”Managing Shared Access to a spectrum Common”, Working Paper Series, ESD-WP-2007-01, Cambridge University, 2006 and 2007.

[3] C. Liu and J. Kaiser J., “A survey of mobile ad-hoc network routing protocols”, University of Magdeburg, Technical representation, 2005.

[4] J. Haillot and F. Guidec F., “A protocol for content-based communication in disconnected mobile ad-hoc networks, Journal of Mobile Information System, vol 6, no2, 2010, pp 123-154.

[5] A. Benchi, P. Launay P. and F. Guidec F., “A JavaSpace Implementation for opportunistic Networks” Proceedings of the International Conference on Future Computational Technologies & Applications, FUTURE COMPUTING, 2012, pp 49-54.

[6] P. Rysavy, “Breakfree with wireless LANs network computing, mobile and wireless technology feature”, 2009, www.rysavy.com/articles/Breakfree.html.

[7] G. Faulhaber, “The Question of Spectrum: Technology, Management, and Regime Change”, paper presented at Michigan University, May 16, 2005.

[8] E. Goodman, “Spectrum Rights in the Telecoms to come”, San Diego Law Review vol 41, 2004, pp 269-404.

[9] E. Friedman, “Fair and Robust Power Allocation Rules for Multiple Access Channels”, Draft, Operation Research & Industrial Engineering, Cornell University, 2005.

[10] A. Odlyzko, “Telecom Dogmas and Spectrum Allocations”, 2004, Wireless unleashed blog www.dtc.umn.edu/doc/network.html

[11] D. Reed, “How Wireless Networks Scale: the Illusion of Spectrum Scarcity”, 2002, Washington, www.jacksournet/spectrumcapacity/FCC.pdf

[12] D. Reed, “Liquid Protocols”, MIT Cambridge Communication , 2005, www.cfp.mit.edu/slides/David_reed_Jun5.pdf.

[13] H. Demstz, “Towards a Theory of Property Rights: competition between private and collective ownership”, Journal of Legal Studies, vol XXX1(2), 2000, pp 653-672.

[14] F. Kelly, “Models for a self-managed Internet” Philosophical Transactions of the Royal Society, 2000, A358.

[15] M. Lepinski and S. Kent, “An infrastructure to support Secure Internet Routing” RFC 6480 February, 2012.

[16] J. Durand, I. Pepelnjuk and G. Doering G., “BGP Operations and Security; Internet Draft, IETF, www.ietf.org/html/draft/bgp.txt retrieved June 30, 2014.

[17] F. J. Ogwu, M. Talib M. and G. A. Aderounmu, “Stochastic Estimator-based Wireless Traffic Control Scheme”, Journal of Computer Science, 3(12), 2007, pp 918-923.

[18] K. Horiokisio, “WiMax Networks and Cognitive Radio research”, 2013, www.cmpe.boun.edu/wico=research.

[19] D. P. Sataparthy, ”An Algorithm for Unlicensed Fixed Power Devices”, IEEE Wireless Communications and Networking Conference (WCNC), September, 2002.

Feedback Optimal Control for Inverted Pendulum Problem by Using the Generating Function Technique

Hany R. Dwidar
Astronomy, Meteorology and Space Science Dept.
Faculty of Science - Cairo University
Giza - EGYPT 12613

***Abstract*—In this paper, a model is described for a system consisting of an inverted pendulum attached to a cart. We design for this model a feedback optimal control based on Linear Quadratic regulator, LQR by using the generating Function technique. This design with hard and soft constraints will help the pendulum to stabilize in the upright position. A solution of the continuous low-thrust optimal control problem based on LQR method is implemented. An example applied to this control design for a hard constraint boundary condition.**

Keywords—Inverted pendulum; Feedback control; Stability analysis

I. Introduction

The traditional problem for the field of control systems is the inverted pendulum system see e.g. [1] , [3] , [5] , [6] and [9]. The system is consist of an inverted pendulum exposed to a torque and attached to a cart which equipped with a motor that drives it along a friction horizontal track. Both the torque and the force produced from the motor of the cart are the feedback-control forces. there are two equilibrium points for the inverted pendulum system, one of them is that when the pendulum is pointing downwards which is stable, the other one is at the upwards position which is unstable. The stable equilibrium requires no control input to be achieved and, thus, is uninteresting from a control perspective. The unstable equilibrium corresponds to a state in which the pendulum points strictly upwards and, thus, requires a control force to maintain this position.

In literature, the feedback control of inverted pendulum control system is made by linearizing the dynamics about the nominal trajectory and by applying the classic control theory to such linear dynamical system. The approach of optimal feedback control using the generating function [8] is very efficient when used to solve the control problem of this system. In order to use the feedback optimal control approach the lateral dynamics is expressed in a state vector form with adding the control forces to the equations of motion.

II. Modelling

In this system, a pendulum with a torque $T(t)$, is attached to the side of cart by means of a pivot which allows the pendulum to swing in the xy-plane. A cart equipped with a motor exerts force $F(t)$, provides horizontal motion of the cart on a friction track, see Fig 1 . The purpose of the torque $T(t)$ and the force $F(t)$ is that keeping the pendulum balanced upright.

A. *Formation of the problem :*

By assuming that the pendulum is a thin rod with length l. Then, applying Newton's second law to the linear and angular displacement, the equations of motion are [4]

$$(M+m)\ddot{x}+\varepsilon\dot{x}+m\frac{l}{2}\ddot{\theta}\cos\theta-m\frac{l}{2}\dot{\theta}^2\sin\theta=F(t) \tag{1}$$

$$m\frac{l}{2}\cos\theta\ddot{x}+\frac{1}{3}ml^2\ddot{\theta}=T(t)+mg\frac{l}{2}\sin\theta \tag{2}$$

where M is the cart mass, m is the pendulum mass, $x(t)$ is displacement of the center of mass of the cart from the center of the inertial frame, $\theta(t)$ is the angle between the pendulum and the top vertical.

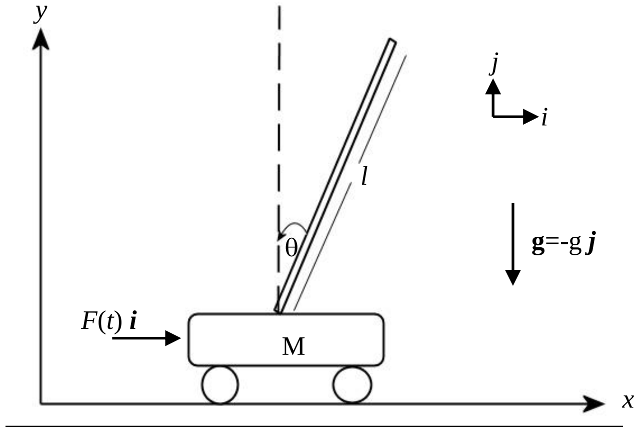


Fig. 1. Inverted Pendulum system

After some calculation, we can obtain

$$\begin{aligned}\ddot{x}=&\frac{4}{3\alpha(\theta)}F(t)-\frac{2\cos\theta}{l\alpha(\theta)}T(t)+\frac{2ml\sin\theta}{3\alpha(\theta)}\dot{\theta}^2-\\&-\frac{4\varepsilon}{3\alpha(\theta)}\dot{x}-\frac{mg\sin2\theta}{2\alpha(\theta)}\end{aligned} \tag{3}$$

$$\begin{aligned}\ddot{\theta}=&\frac{4(M+m)}{ml^2\alpha(\theta)}T(t)-\frac{2\cos\theta}{l\alpha(\theta)}F(t)-\frac{m\sin2\theta}{2\alpha(\theta)}\dot{\theta}^2+\\&+\frac{2\varepsilon\cos\theta}{l\alpha(\theta)}\dot{x}+\frac{2(M+m)g\sin\theta}{l\alpha(\theta)}\end{aligned} \tag{4}$$

where

$$\alpha(\theta) = \frac{4}{3}(M+m) - m\cos^2\theta = \frac{1}{3}(4M+m) + m\sin^2\theta \tag{5}$$

and g is the acceleration due to gravity and equal 9.8 m/s.

By introducing the following variables for a more convenient form of the equations (3) and (4)

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4)^T = (x, \theta, \dot{x}, \dot{\theta})^T \tag{6}$$

We obtain the following equivalent first-order system

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{y}) = \begin{bmatrix} y_3 \\ y_4 \\ \frac{1}{\alpha(y_2)}\left(\frac{4}{3}F(t) - \frac{2\cos y_2}{l}T(t) - \frac{4}{3}\varepsilon y_3 + \frac{2ml\sin y_2}{3}y_4^2 - \frac{mg\sin 2y_2}{2}\right) \\ \frac{1}{\alpha(y_2)}\left(\frac{4(M+m)}{ml^2}T(t) - \frac{2\cos y_2}{l}F(t) - \frac{m\sin 2y_2}{2}y_4^2 + \frac{2\varepsilon\cos y_2}{l}y_3 + \frac{2(M+m)g\sin y_2}{l}\right) \end{bmatrix} \tag{7}$$

where

$$\alpha(y_2) = \frac{1}{3}(4M+m) + m\sin^2 y_2 \tag{8}$$

It is well known that for (7) with no control ($F(t)$ = 0 and $T(t)$=0), the cart at rest with the pendulum in the upright position is an unstable equilibrium, while the cart at rest with the pendulum in the downward position is a stable equilibrium. Our concern is that when the pendulum at the unstable equilibrium point, so in the following section the system will be linearized about the unstable equilibrium $(0,0,0,0)^T$.

B. Linearization:

Now, by putting (7) in the following form

$$\dot{\boldsymbol{y}} = C(\boldsymbol{y})\boldsymbol{y} + D(\boldsymbol{y})\boldsymbol{u} \tag{9}$$

where

$$C(\boldsymbol{y}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{mg}{\alpha(y_2)}\frac{\sin 2y_2}{2y_2} & -\frac{4\varepsilon}{3\alpha(y_2)} & \frac{2ml\sin y_2}{3\alpha(y_2)}y_4 \\ 0 & \frac{2(M+m)g}{l\alpha(y_2)}\frac{\sin y_2}{y_2} & \frac{2\varepsilon\cos y_2}{l\alpha(y_2)} & -\frac{m\sin 2y_2}{2\alpha(y_2)}y_4 \end{bmatrix} \tag{10}$$

$$D(\boldsymbol{y}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{4}{3\alpha(y_2)} & -\frac{2\cos y_2}{l\alpha(y_2)} \\ -\frac{2\cos y_2}{l\alpha(y_2)} & \frac{4(M+m)}{ml^2\alpha(y_2)} \end{bmatrix}, \tag{11}$$

and

$$\boldsymbol{u} = \begin{bmatrix} F(t) & T(t) \end{bmatrix}^T \tag{12}$$

The system (9) will be linearized about the nominal trajectory $(0,0,0,0)^T$. Now by applying that $\lim_{y_2 \to 0}\frac{\sin y_2}{y_2} = 1$ and $\lim_{y_2 \to 0}\cos y_2 = 1$, we can deduce that

$$\dot{\boldsymbol{y}} = A\boldsymbol{y} + B\boldsymbol{u} \tag{13}$$

where

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{3mg}{4M+m} & -\frac{4\varepsilon}{4M+m} & 0 \\ 0 & \frac{6(M+m)g}{l(4M+m)} & \frac{6\varepsilon}{l(4M+m)} & 0 \end{bmatrix} \tag{14}$$

and

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{4}{4M+m} & -\frac{6}{l(4M+m)} \\ -\frac{6}{l(4M+m)} & \frac{12(M+m)}{ml^2(4M+m)} \end{bmatrix} \tag{15}$$

This linearized system (13) allow to apply the feedback optimal control to balance the inverted pendulum around the point of linearization, as seeing in the following section.

III. Closed Loop (Feedback) Optimal Control Problem

The feedback optimal control problem is introduced to find optimal solutions minimize a certain performance index starting from a generic initial state y_0, The outcome is a control law written in terms of the time and the initial state, $v = v(y_0, t_0, t)$, $t_0 \le t \le t_f$.

This represents a closed loop solution: given any initial state y_0 at the time t_0, it is possible to evaluate the optimal solution starting from such state up to the final target. If for any reason the state is perturbed and assumes a new value $y'_0 = y_0 + \delta x,\ t'_0 = t_0 + \delta t$, we are able to compute the new optimal solution by simply evaluating $v = v(y'_0, t_0, t)$, avoiding, in this way, the solution of another two-point boundary value problem. Thus, a trajectory designed in this way has the property to respond to errors that occur during the transfer. Another important aspect of this approach is the robustness of the solution. Once the optimal feedback control problem is solved, the solution $v = v(y_0, t_0, t)$ is available. Analyzing this function, the control law that is less sensitive to changes in the initial conditions can be chosen as nominal solution. This solution is said to be robust with respect to the initial conditions.

A. *Solving the Feedback Optimal Linear Quadratic Terminal Controller Using the Generating Function Technique*

Consider the problem of minimizing the following performance index , [7] , [8]

$$J = \frac{1}{2} y_f^T Q_f y_f + \frac{1}{2}\int_{t_0}^{t_f} (y^T Q y + v^T R v) dt$$

subject to the linear dynamics

$$y = Ay + Bv \tag{16}$$

and with the given initial and final conditions

$$y(t_0) = y_0 \qquad \text{and} \qquad y(t_f) = y_f$$

According to the classical theory, the Hamiltonian of the optimal control problem is

$$H = \frac{1}{2}(y^T Q y + v^T R v) + \lambda^T (Ay + Bv) \tag{17}$$

where the set of Lagrangian multipliers has been introduced. From the optimality condition

$$\frac{\partial H}{\partial v} = 0 \tag{18}$$

It is possible to get an explicit expression for the control in terms of the Lagrangian multipliers

$$v = -R^{-1} B^T \lambda \tag{19}$$

Substituting the expression of given by equation (19), the Hamiltonian (17) turns out to be

$$H = \frac{1}{2}\begin{pmatrix} y \\ \lambda \end{pmatrix}^T \begin{bmatrix} Q & A^T \\ A & -BR^{-1}B^T \end{bmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} \tag{20}$$

while the dynamics of the system and that of the Lagrange multipliers reduces to

$$\begin{pmatrix} \dot{y} \\ \dot{\lambda} \end{pmatrix} = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \begin{pmatrix} y \\ \lambda \end{pmatrix} \tag{21}$$

Suppose now that we have a generating function $F_2(y, \lambda_0, t, t_0)$ for the transformation between a fixed state (y_0, λ_0, t_0) and a moving state (y, λ, t_0) . This transformation is canonical because it preserves the area in the phase space and in addition generates the identity transformation at $t = t_0$. we can derive this generating functions and their associated relations for this canonical transformation from Hamilton-Jacobi PDE

$$\lambda = \frac{\partial F_2(y, \lambda_0, t, t_0)}{\partial y} \tag{22.a}$$

$$y_0 = \frac{\partial F_2(y, \lambda_0, t, t_0)}{\partial \lambda_0} \tag{22.b}$$

$$0 = \frac{\partial F_2(y, \lambda_0, t, t_0)}{\partial t} + H\left(y, \frac{\partial F_2(y, \lambda_0, t, t_0)}{\partial y}, t \right) \tag{22.c}$$

Since the Hamiltonian is quadratic, F_2 can be put in a quadratic form as follows

$$F_2(y, \lambda_0, t, t_0) = \frac{1}{2}\begin{pmatrix} y \\ \lambda_0 \end{pmatrix}^T \begin{bmatrix} F_{yy}(t, t_0) & F_{y\lambda_0}(t, t_0) \\ F_{\lambda_0 y}(t, t_0) & F_{\lambda_0\lambda_0}(t, t_0) \end{bmatrix} \begin{pmatrix} y \\ \lambda_0 \end{pmatrix} \tag{23}$$

which can be used to find the unknown boundary conditions using the given ones. From the properties of F_2 we have

$$\lambda = \frac{\partial F_2}{\partial y} = \begin{pmatrix} F_{yy} & F_{y\lambda_0} \end{pmatrix} \begin{pmatrix} y \\ \lambda_0 \end{pmatrix} \tag{24}$$

The Hamiltonian (20) can be expressed as a function of (y, λ_0) by using equation (21)

$$H = \frac{1}{2}\begin{pmatrix} y \\ \lambda_0 \end{pmatrix}^T \begin{bmatrix} I & F_{yy} \\ 0 & F_{\lambda_0 y} \end{bmatrix} \begin{bmatrix} Q & A^T \\ A & -BR^{-1}B^T \end{bmatrix} \begin{bmatrix} I & 0 \\ F_{yy} & F_{y\lambda_0} \end{bmatrix} \begin{pmatrix} y \\ \lambda_0 \end{pmatrix} \tag{25}$$

Since the Hamiltonian at the fixed state can be taken zero without any loss of generality, then the Hamiltonian of the moving state and the generating function satisfy the Hamilton-Jacobi PDE (22.c)

$$0 = \begin{pmatrix} y \\ \lambda_0 \end{pmatrix}^T \left\{ \begin{bmatrix} \dot{F}_{yy} & \dot{F}_{y\lambda_0} \\ \dot{F}_{\lambda_0 y} & \dot{F}_{\lambda_0\lambda_0} \end{bmatrix} + \right. \\ \left. + \begin{bmatrix} I & F_{yy} \\ 0 & F_{\lambda_0 y} \end{bmatrix} \begin{bmatrix} Q & A^T \\ A & -BR^{-1}B^T \end{bmatrix} \begin{bmatrix} I & 0 \\ F_{yy} & F_{y\lambda_0} \end{bmatrix} \right\} \begin{pmatrix} y \\ \lambda_0 \end{pmatrix} , \tag{26}$$

whose sub-matrix components provide the following set of matrix ODEs for (Riccati Equations [2]) $F_{yy}(t, t_0)$, $F_{y\lambda_0}(t, t_0) = F^T_{\lambda_0 y}(t, t_0)$, and $F_{\lambda_0\lambda_0}(t, t_0)$

$$\dot{F}_{yy} + Q + F_{yy}A + A^T F_{yy} - F_{yy}BR^{-1}B^T F_{yy} = 0 ,$$

$$\dot{F}_{y\lambda_0} + A^T F_{y\lambda_0} - F_{yy}BR^{-1}B^T F_{y\lambda_0} = 0 , \tag{27}$$

$$\dot{F}_{\lambda_0\lambda_0} - F_{\lambda_0 y} B R^{-1} B^T F_{y\lambda_0} = 0 \ ,$$

The initial conditions which verify the identity transformation at $t = t_0$ are

$$F_{yy}(t_0, t_0) = 0_{n\times n} \ ,$$
$$F_{y\lambda_0}(t_0, t_0) = I_{n\times n} \ , \qquad (28)$$
$$F_{\lambda_0\lambda_0}(t_0, t_0) = 0_{n\times n}$$

B. *Computing the Generating function*

- The HCP

We compute $F_2(y, \lambda_0, t, t_f)$ by using t_f as our initial time, then we have from (23)

$$y_0 = \frac{\partial F_2(y, \lambda_0, t, t_0)}{\partial \lambda_0} = F_{\lambda_0 y} y_f + F_{\lambda_0\lambda_0}\lambda_0 \qquad (29)$$

Since we have y_0 and y_f are given the initial Lagrange multiplier can be evaluated through

$$\lambda_0 = F^{-1}_{\lambda_0\lambda_0}(t_f, t_0)(y_0 - F_{\lambda_0 y}(t_f, t_0) y_f) \qquad (30)$$

Using (30) we can get the optimal trajectory by forward integration of (21). Since this relation (30) is valid for any initial time $t \le t_f$, we have

$$\lambda(t) = F^{-1}_{\lambda_0\lambda_0}(t_f, t)(y(t) - F_{\lambda_0 y}(t_f, t) y_f)$$

and therefore the control can be given by from (21)

$$v(t) = -R^{-1} B^T F^{-1}_{\lambda_0\lambda_0}(t_f, t)(y(t) - F_{\lambda_0 y}(t_f, t) y_f)$$

- The SCP

We compute F_1 from Legendre transformation as follows

$$F_1(y, y_0, t, t_0) = F_2(y, \lambda_0, t, t_0) - y_0^T \lambda_0 \ ,$$

Substituting y_0 from (29) we get after some algebraic manipulations

$$F_1(y, y_0, t, t_0) = \frac{1}{2}\begin{pmatrix} y \\ y_0 \end{pmatrix}^T \begin{bmatrix} F_{yy} - F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y} & F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} \\ F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y} & -F^{-1}_{\lambda_0\lambda_0} \end{bmatrix} \begin{pmatrix} y \\ y_0 \end{pmatrix} \qquad (31)$$

then we have

$$\lambda_0 = F^{-1}_{\lambda_0\lambda_0}(t_f, t_0)(y_0 - F_{\lambda_0 y}(t_f, t_0) y_f) \ ,$$
$$\lambda_f = F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0}(t_f, t_0) y_0 + (F_{yy} - F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y})(t_f, t_0) y_f \ ,$$
$$\lambda_f = Q_f y_f \ ,$$

By equating the second and the third equations we get a relation between y_0 and y_f which can be plugged into the first equation to get

$$\lambda_0 = \Big\{\Big[F^{-1}_{\lambda_0\lambda_0} - \\ - F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\left(Q_f - F_{yy} + F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\right)^{-1} F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0}\Big](t_f, t_0)\Big\} y_0 \ , \qquad (32)$$

Using (32) we can get the optimal trajectory by forward integration of (21). Since this relation (32) is valid for any initial time $t \le t_f$, we have

$$\lambda(t) = \Big\{\Big[F^{-1}_{\lambda_0\lambda_0} - \\ - F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\left(Q_f - F_{yy} + F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\right)^{-1} F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0}\Big](t_f, t_0)\Big\} y(t),$$

and the control is given by

$$v(t) = -R^{-1} B^T \Big\{\Big[F^{-1}_{\lambda_0\lambda_0} - \\ - F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\left(Q_f - F_{yy} + F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0} F_{\lambda_0 y}\right)^{-1} F_{y\lambda_0} F^{-1}_{\lambda_0\lambda_0}\Big](t_f, t_0)\Big\} y(t) \ , \qquad (33)$$

When Q_f is large enough the SCP solution converges to the HCP one.

IV. Numerical Example

The values of the parameters for our numerical example are given as follow,

$M = 0.8 kg$, $m = 0.21\, kg$ and $\varepsilon = 2.1$.

and

$t_0 = 0$ and $\boldsymbol{y}_0 = \left(0.001 \quad 0.001 \quad 0.0002 \quad 0.0001\right)^T$,

$t_f = 0.45\ s$ and $\boldsymbol{y}_f = \left(0.0 \quad 0.0 \quad 0.0 \quad 0.0\right)^T$.

Then, by substituting in (14) and (15) we have

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1.8106 & -2.4633 & 0 \\ 0 & 28.5506 & 6.0574 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1.1730 & -2.8845 \\ -2.8845 & 45.4851 \end{bmatrix}.$$

and by putting

$$Q = \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$Q_f = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We plot the resulting data in the following figures

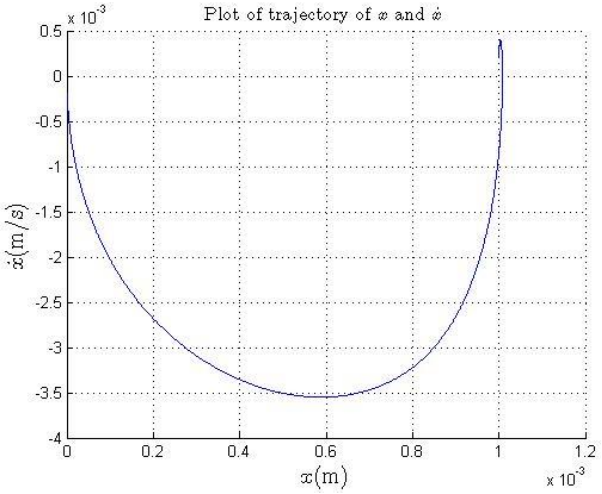


Fig. 2. Plot of trajectory of x and $\dot{\boldsymbol{x}}$

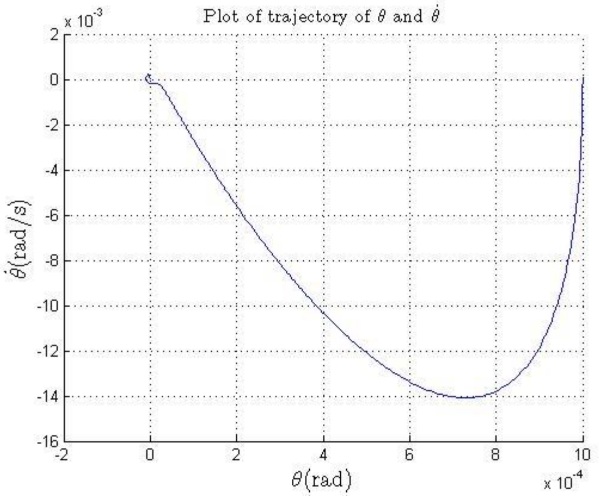


Fig. 3. Plot of trajectory of $\boldsymbol{\theta}$ and $\dot{\boldsymbol{\theta}}$

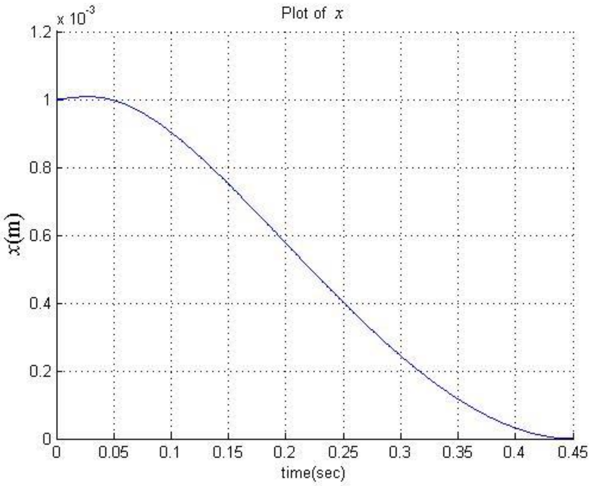


Fig. 4. Plot of $x(t)$

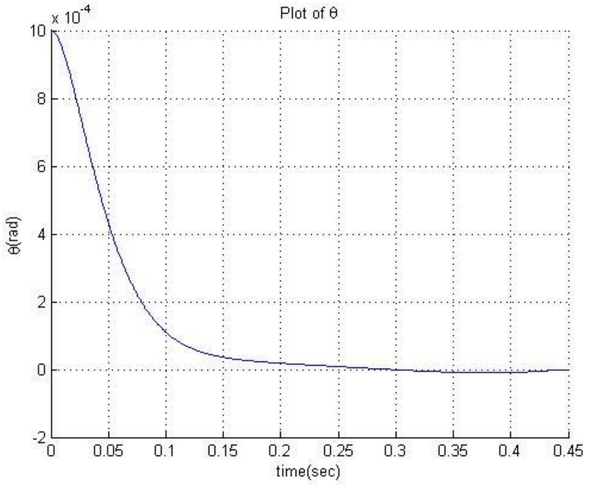


Fig. 5. Plot of $\theta(t)$

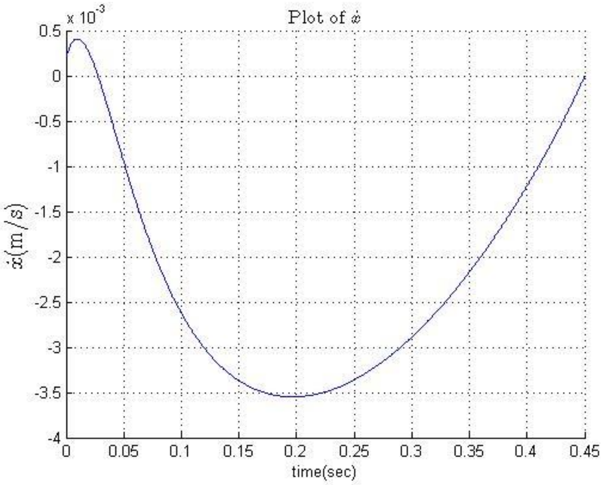


Fig. 6. Plot of $\dot{x}(t)$

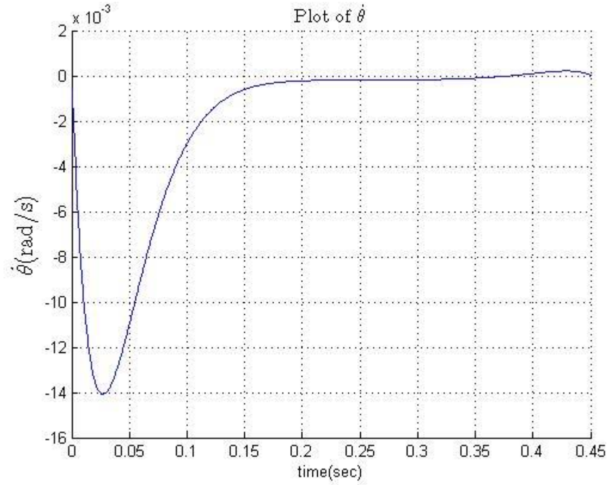


Fig. 7. Plot of $\dot{\theta}(t)$

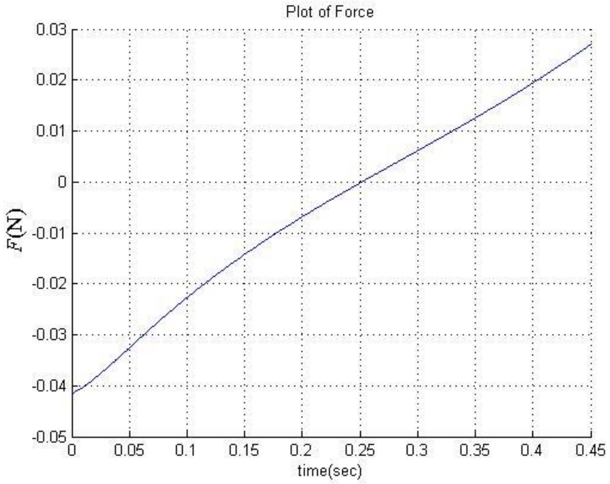


Fig. 8. Plot of $F(t)$

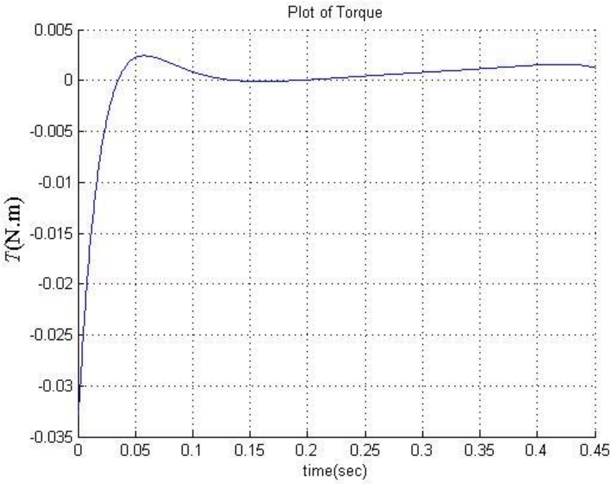


Fig. 9. Plot of $T(t)$

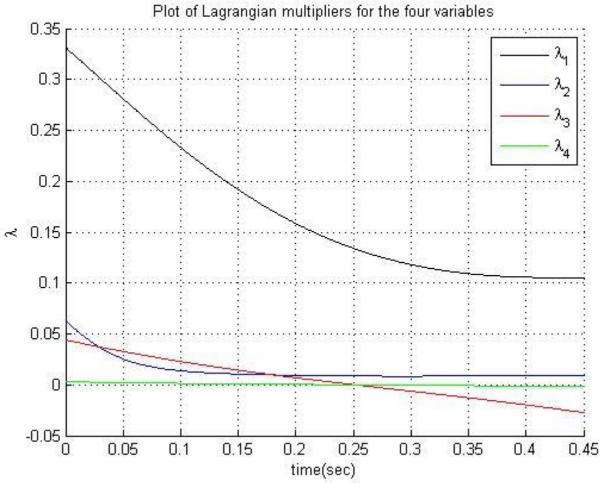


Fig. 10. Plot of Lagrangian multiliers

V. Conclusion

In the present study, the feedback optimal control of highly nonlinear inverted pendulum problem is solved by linearizing the original nonlinear dynamics. The linearized problem has been solved using the generating function technique where the method can be used for both hard and soft constraint boundary condition.

A proof exists in the very simple case of linear hard constraint problem (the final state is fully specified) and the figures (Fig2 to Fig 10) for the numerical example showed an excellent result in keeping the inverted pendulum in the unstable state for a short time. For a future work we can apply the same method feedback optimal control by using the generating function technique on the double inverted pendulum.

References

[1] F. Atay, "Balancing the inverted pendulum using position feedback," Applied Mathematics Letter, vol. 12, pp. 51-56, 1999.

[2] A. E. Bryson and Y. Ho, Applied Optimal Control, London, England: Hemisphere Publishing Corp., 1975.

[3] K. Furuta, M. Yamakita and S. Kobayashi, "Swing-up control of inverted pendulum using pseudo-state feedback," Journal of Systems and Control Engineering, vol. 206, pp. 263-269, 1992.

[4] M. Landry, S. A. Campbell, K. Morris and C. O. Aguilar, "Dynamics of an Inverted Pendulum with Delayed Feedback Control," SIAM J. APPLIED DYNAMICAL SYSTEM, vol. 4, no. 2, pp. 333-351, 2005.

[5] R. Lozano and I. Fantoni, "Passivity based control of the inverted pendulums. In Normand-Cyrot," in Perspectives in control, 1998.

[6] H. Meier, Z. Farwig and H. Unbehauen, "Discrete computer control of a triple-inverted pendulum," Optical Control Applications and Methods,, vol. 11, pp. 157-171, 1990.

[7] C. Park and D. J. Scheeres, "Determination of optimal feedback terminal controllers for general boundary conditions using generating functions," Automatica, vol. 42, p. 869 – 875, 2006.

[8] C. Park and D. J. Scheeres, "Solutions of Optimal Feedback Control Problems with General Boundary Conditions using Hamiltonian Dynamics and Generating Functions," in Proceeding of the American Control Conference, Boston, Massachusetts, 2004.

[9] J. Sieber and B. Krauskopf, "Bifurcation analysis of an inverted pendulum with delayed feedback control near a triple-zero eigenvalue singularity," Nonlinearity, vol. 17, pp. 85-103, 2004.

Efficient Identification of Common Subsequences from Big Data Streams Using Sliding Window Technique

Adi Alhudhaif
Department of Computer Science, The George Washington University, Washington, DC. 20052, USA
Department of Computer Science, Salman Bin Abdulaziz University,
Al Kharj, The Kingdom of Saudi Arabia

Abstract—**We propose an efficient Frequent Sequence Stream algorithm for identifying the top k most frequent subsequences over big data streams. Our Sequence Stream algorithm gains its efficiency by its time complexity of linear time and very limited space complexity. With a pre-specified subsequence window size S and the k value, in very high probabilities, the Sequence Stream algorithm retrieve the top k most frequent subsequences of size S. The Stream Sequence algorithm also provides a high accuracy of the estimation of the number of occurrences of each promoted subsequence. Our experiments indicate several factors that influence the result accuracy of the Sequence Stream algorithm: stream size, subsequence size S and frequency of the subsequence.**

Keywords—Frequent subsequence; Stream processing; Periodic pattern; Pattern recognition; Big data processing

I. Introduction

Due to the new data collection methods, vast amount of data is produced [1]. This data-increasing trend is associated with business needs, geographical research works, social media networks and etc. and these result in “Big Data”. Big Data situation relates to the problem of dealing with very large amounts of data [2]. It presents a qualitatively different state of affairs for the organization of information processing, namely, this organization cannot utilize all the data explicitly. Data processing is one of the important challenges and many studies have been made on this topic [3]. Big Data support to build several worldwide social network connections, which integrate human beings with the accelerated progress of communication. Because of big data, entrepreneurs could make wise decision based on consumers’ behaviors. Recently, the use of big data has a key role in improving competitiveness in all kinds of fields. The big data stream contains very large amounts of information. The stream data processing is to understand data and to retrieve useful information from it. Various methods are designed to deal with big data [4][5][6]. The challenges include accuracy, efficiency and availability.

Frequent sequence mining finds sets of data elements that occur together frequently in many subsequences. Frequent sequence mining, which retrieve the most frequent subsequences from a stream of a very large sequence. It gained a great deal of attention in the field of data mining due to its great value in many applications, such as: trend prediction, stock market, DNA sequence analysis (Bioinformatics), using history of side affects or symptoms to predict valuable medical information, web user analysis, finding language or linguistic sequences from natural language text.

In this paper, we introduce a novel technique for mining the top k frequent subsequences over large stream of big data with a pre-defined subsequence size S, in the fashion of stream processing. The algorithm provides very high probabilities for retrieving the most frequent subsequences in leaner time using very limited space and memory locations.

II. Frequent Sequence Mining in Stream Processing

Finding most frequent sequences is considered as one of the most heavily studied data mining task since its introduction in work [7] and is of wide scientific interest [8][9][10][11]. Subsequences are valuable kind of data that occur more often in domains such as: information security, artificial intelligent, machine learning, education, medical, financial and many other fields. As for medical field, extracting frequent subsequences from very large DNA sequences is a key step for understanding biological processes as basic as the RNA transcription [12].

Stream processing uses different methods comparing to traditional datasets computing, it requires relatively smaller respond time with dealing huge amount of data. In computer science, the streaming algorithms are designing for processing data streams in the way of limited time and limited memory. It was first introduced in 1999 [13] [14], and then spread to all kinds of angles in computer science, such as database, networking and machine learning. Now the big data society comes to study stream algorithms when large amounts of data can be operated continuously regardless of storage and access distribution, meanwhile respond quickly to new information. In reality, stock market data is a typical stream data. The data contains real-time price, transaction and other financial information. Traders usually receive and analyze data streams to make decisions by advanced systems.

We focus on the process of massive stream by optimal processing algorithm to extract meaningful value from large sequence of big data. This is done by retrieving (on-the-fly [15][16]) the most frequent subsequences over large stream of big data with the concerns of time-consumption and space-consumption.

III. Frequent Sequence Stream Algorithm

The Frequent Sequence Stream algorithm (FSS), was inspired during the development of Multi-Buffer based algorithm in work [5]. Multi-Buffer based algorithm was proposed to extract the top k most frequent elements over large stream of Big Data. FSS works in window sliding technique and window size is a pre-defined value of S. In addition, FSS holds multi sequence candidates (SeqCan) that hold common sequences of size S. For a Big data stream $u = u_1,...,u_n \in \Sigma^*$ we consider all the subsequences $u_iu_{i+1} ...u_{i+S-1} \sqsubseteq u$, where $1 \leqslant S \leqslant n$ and $1 \leqslant i_1 < i_2 < \cdots < i_n \leqslant n$, and goal is to find the top k most frequent sequences of size S and an approximate counter that reflects each subsequence occurrence.

Using k Sequence candidates, the FSS algorithm can be stated as following: store the first new arrival of sequence $(u_iu_{i+1} ...u_{i+S-1})$ to SeqCan#1 and set Weight (w) to β and Counter (c) to 1. Keep comparing the incoming new sequence with the previously stored sequences candidates (SeqCan). If the new incoming new sequence equals to one of the sequence candidates in (SeqCan), increase its associated counter by 1 and increase its weight by β.

Otherwise, assign this new subsequence to any sequence candidates that has an associated weight equal to zero and set that weight to β and its counter to 1. At the case of no weight equals to zero, decrease the weight (w) with minimum value by 1. By the end of this stream, output stored sequences (Candidates) and their associated counters (c).

For example; when k = 3 the algorithm FSS can be described as follows:

```
Repeat
        Get next sequence using sliding window of size S
        Seq = $u_iu_{i+1} ...u_{i+S-1}$
        if ( w1 ≠ 0 and SeqCan1 = Seq ):
                w1=w1+ β, c1=c1+1
        Else_if ( w2 ≠ 0 and SeqCan2 = Seq):
                w2=w2+ β, c2=c2+1
        Else_if ( w3 ≠ 0 and SeqCan3 = Seq):
                w3=w3+ β, c3=c3+1
        Else _if (SeqCan1 ≠ Seq and w1 ≠ 0) and (SeqCan2 ≠ Seq
        and w2 ≠ 0) and (SeqCan3 ≠ Seq and w3 ≠ 0):
             Minimum [w1,w2,w3] =  Minimum[w1,w2,w3] - 1
        Else_if (w1 = 0 ):
                w1 = β , c1 = 1 , CanSeq1 = Seq
        Else_if (w2 = 0 ):
                w2 = β , c2 = 1 , CanSeq2 = Seq
        Else_if (w1 = 0 ):
                w3 = β , c3 = 1 , CanSeq3 = Seq
        Move window by 1( i = i + 1).
Until no more sequences
```

Moreover, the output of the FSS algorithm will be k pairs (candidate, counter). The focus of this algorithm is to improve the probability that one of the k pairs contains the most frequent sequence of size S, and enhance the accuracy of estimating its frequency. The FSS algorithm is able to select up to k – 1 top frequent sequences in the data stream. For example, when k = 3 and an input of random sequences with two top occurrence of frequency 12% and 15%, they would be selected efficiently by using three sequence candidates (SeqCan) or more.

IV. Experimints

For every single stream file with determined sequence frequncy we generated many iterations using The Fisher-Yates shuffle algorithm [17][18]. Generating pseudo-random numbers was done using both generator functions in Python's library Lib/random.py and the random number library in C that takes variable seeds such as: current system time to generate pseudo-random numbers. Then, according to the most frequent frequence.

We performed and examined Frequent Sequence Stream (FSS) algorithm using the big data stream under a common implementation framework to test their performance as accurately as possible. The algorithm was implemented using both C and Python, and compiled using gcc on Cygwin 1.7.25 for C code, and Python 2.7.5 for python code. We ran Python experiments on 2.6GHz dual-core Intel Core i5 with 8GB of RAM running OS X 10.9.2. Experiments of algorithms in C were ran on Intel 4th generation core i5 using 8GB of RAM running Microsoft Windows Server 2012. We did not observe notewothy differences between two compilers.

A. The Calculation of Sequence Frequency

In big data stream of size *n*, and a pre-defined subsequence size S. A subsequence X has a frequency of 100% when the number of occurrence of sequence X is $\lfloor n/S \rfloor$. For example: stream of size 100,000 elements and a subsequence size of 7, the subsequence X has a frequency of 15% when it occurs 2,142 times ($\lfloor (n/S)* 0.15 \rfloor$).

B. Results

Using stream sizes 30,000, 100,000 and 1,000,000 Fig. 1 shows the probabilities of retrieving the most frequent subsequences of size 3, with low frequencies: 5%, 4%, 3%, 2% and 1%.

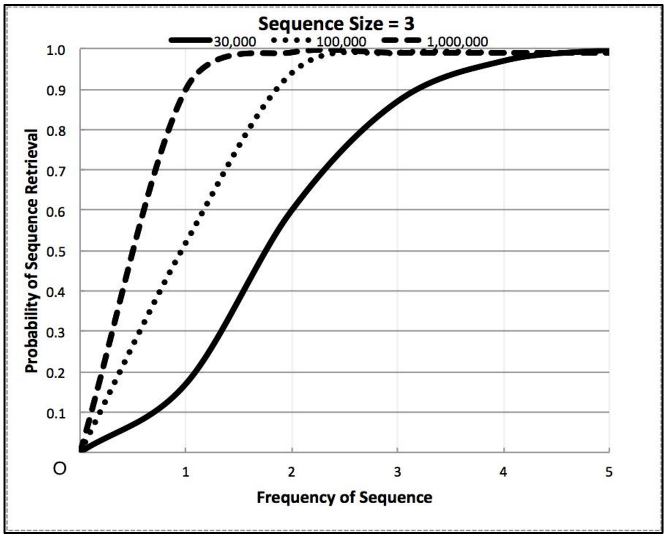


Fig. 1. Probabilities of extracting most frequent subsequences of size 3.

Fig. 2 represents the probabilities for retrieving most frequent subsequences of size 7 with low frequencies using various stream sizes.

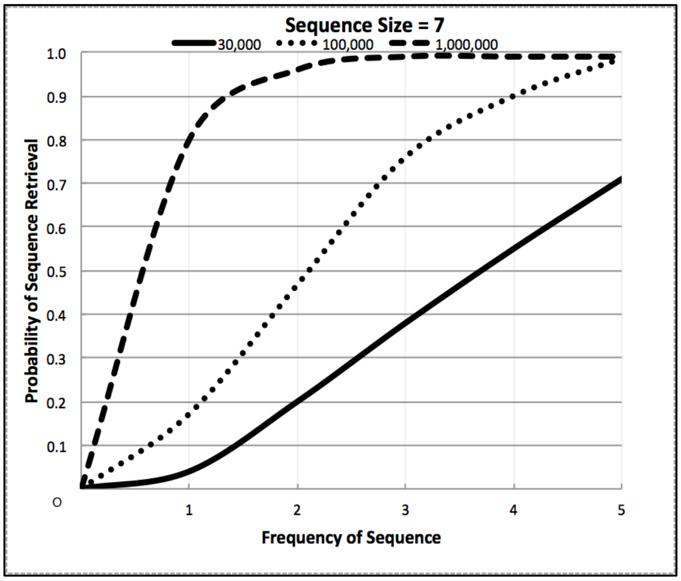


Fig. 2. Probabilities of extracting subsequences of size 7 using various stream sizes

Moreover, in Fig. 3 we increased the subsequence size to be matched to 15 using various stream sizes: 30,000, 100,000 and 1,000,000.

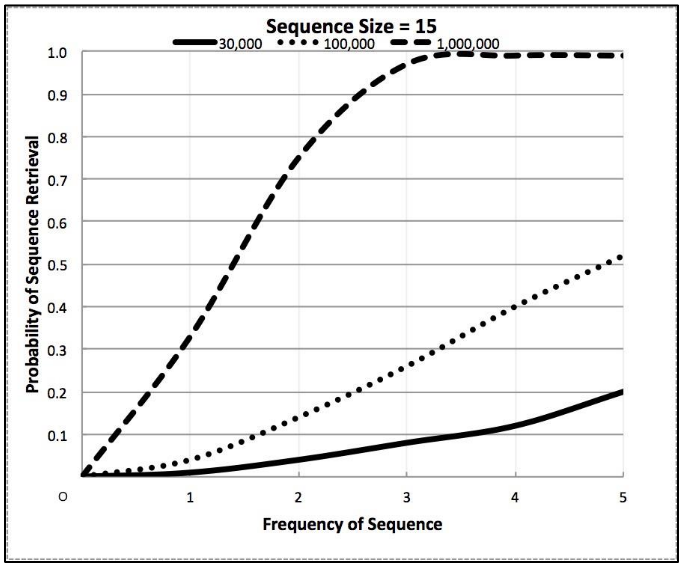


Fig. 3. Probabilities for extracting subsequences of size 15 within various stream sizes.

For larger subsequence frequencies like 10%, Fig. 4 shows the probabilities for retrieving subsequences of various sizes versus various stream sizes.

Counters that represent the number of a subsequence occurrence are part of the FSS algorithm. For a stream size of 100,000, and subsequences of low frequencies, Fig. 5 shows the accuracy of counter values returned by the FSS algorithm compared to the real number of occurrences within the big data stream.

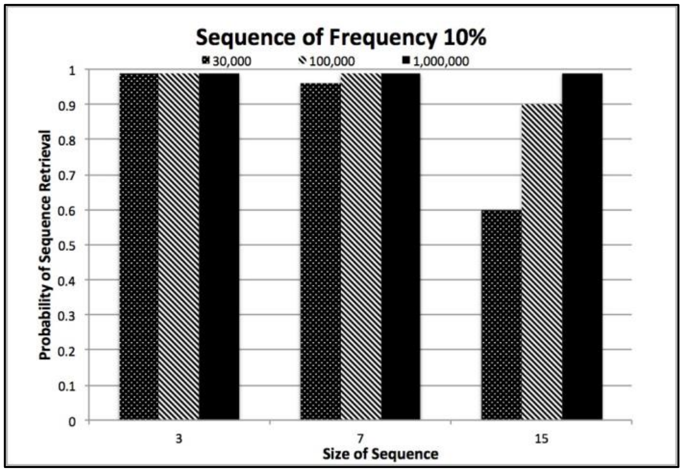


Fig. 4. Probabilities for retrieving subsequences of various stream and subsequences sizes

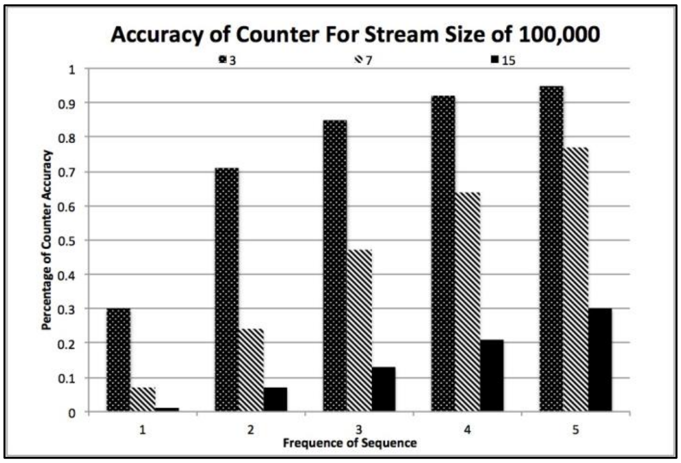


Fig. 5. Accuracy of counter values using various subsequences sizes and stream size of 100,000.

For a stream size of 1,000,000 and subsequences of low frequencies, Fig. 6 shows the accuracy of counter values returned by the FSS algorithm compared to the real number of occurrences.

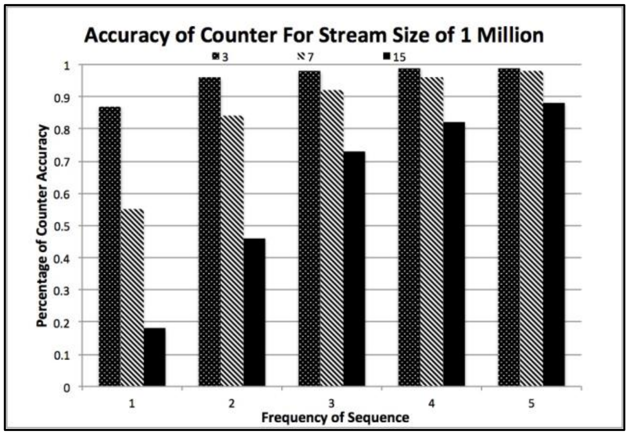


Fig. 6. Accuracy of counter values using various subsequences sizes and stream size of 100,000.

V. Comments/Future Research

During experiments many values for β were tested and verified. The best value for β turned out to be S^2. As for Multi-Buffer based algorithm in work [5], stream sizes has an impact on the performance of the FSS algorithm as shown in Fig. 1, Fig. 2 and Fig. 3. The bigger the stream size the better the results.

Moreover, we observed that the subsequence sizes have an influence on the output results of the FSS algorithm. The smaller subsequences' sizes to be processed, the better accuracy of the results to be promoted. To explain this, for a stream size of 100,000, 1% frequency of a subsequence size 3 is 333 times, 1% frequency of a subsequence size 7 is 142 times, and a 1% frequency of subsequence size 15 is 66 times.

By tracking changes of these sequence candidates and their associated weights, we find that the entire process can be divided into two important stages: stable stage and unstable stage explained in work [5]. For subsequences of high frequencies such as %10 and more, probabilities for retrieving those frequencies are very high compared to low frequencies.

As for counter values, Fig. 5 and Fig. 6 show that the stream size factor influence the accuracy of the counter value returned by the FSS algorithm. Counter accuracy increases when the stream size increases. In addition, frequencies of subsequence also impact the accuracy of the returned counter vales.

Above experiments were performed using three sequence candidates (k = 3), we observed a minor enhancement when using more sequences candidates. Moreover, more factors (range of data, number of candidates, number of frequent subsequences) could impact the accuracy of results but this needs to be verified by more experiments. Also, the optimal value of β, its relationship with subsequence's size and subsequence's frequency worth more investigations, and whither has β and number of top frequent subsequences are related.

Acknowledgment

Many thanks for Professor Simon Y Berkovich of Computer Science Department of George Washington University, for his guidance and support.

References

[1] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. 2011.

[2] Simon Berkovich, "Physical World as an Internet of Things" COM.Geo '11: Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications, May 2011, p.66

[3] Berkovich, S., Liao, D.: On Clusterization of Big Data Streams. In: 3rd International Conferenceon Computing for Geospatial Research and Applications, article no. 26. ACM Press,New York (2012)

[4] Zikopoulos P, Eaton C. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[5] Adi Alhudhaif, Tong Yan and Simon Berkovich. "On the organization of cluster voting with massive distributed streams", in Proceedings of the 5th international Conference on Computing for Geospatial Research & Application, Washington, D.C., 2014.COM.Geo

[6] Adi Alhudhaif, Tong Yan and Simon Berkovich, "A cyber-physical algorithm for selecting a prevalent element from big data streams", GSTF Journal on Computing (JoC) Vol 4 No 1.

[7] R. Agrawal and R. Srikant. Mining sequential patterns. In ICDE'95, pages 3–14.

[8] Eric P. Xing, Michael I. Jordan, Richard M. Karp, and Stuart Russell. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. In In Proc. of Advances in Neural Information Processing Systems, pages 200–3. MIT Press, 2003

[9] Pavel A. Pevzner and Sing-Hoi Sze. Combinatorial approaches to finding subtle signals in dna sequences. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pages 269–278. AAAI Press, 2000.

[10] Jean-Marc Fellous, Paul H. E. Tiesinga, Peter J. Thomas, and Terrence J. Sejnowski. Discovering Spike Patterns in Neuronal Responses. J. Neurosci., 24(12):2989–3001, 2004.

[11] Nebojsa Jojic, Vladimir Jojic, Brendan Frey, Christopher Meek, and David Heckerman. Using "epitomes" to model genetic diversity: Rational design of HIV vaccine cocktails. In Y. Weiss, B. Sch¨olkopf, and J. Platt, editors, Advances in Neural Information Pro- cessing Systems 18, pages 587–594. MIT Press, Cambridge, MA, 2006.

[12] P.P. Kuksa and V. Pavlovic, "Efficient discovery of common patterns in sequences over large alphabets", in DIMACS Technical Report, 2009.

[13] Alon, Noga, Yossi Matias, and Mario Szegedy. "The space complexity of approximating the frequency moments." Proceedings of the 28th annual ACM symposium on Theory of computing. ACM, 1996

[14] Babcock, Brian; Babu, Shivnath; Datar, Mayur; Motwani, Rajeev; Widom, Jennifer (2002), "Models and issues in data stream systems", Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002), pp. 1–16, doi:10.1145/543613.543615.

[15] Duoduo Liao, "Real-Time Solid Voxelization Using Multi-Core Pipelining", The George Washington University, February 2009 http://gradworks.umi.com/3344878.pdf.

[16] Duoduo Liao and Simon Y. Berkovich, "A New Multi-Core Pipelined Architecture for Executing Sequential Programs for Parallel Geospatial Computing", in Proceedings of the 1st international Conference on Computing for Geospatial Research & Application, Washington, D.C., June 21 - 23, 2010. COM.Geo '10, ACM, New York, NY, U.S.A., 2010.

[17] Richard Durstenfeld, Algorithm 235: Random permutation, Communications of the ACM, v.7 n.7, p.420, July 1964.

[18] Fisher, Ronald A.; Yates, Frank (1948) [1938]. Statistical tables for biological, agricultural and medical research (3rd ed.). London: Oliver & Boyd. pp. 26–27.

A Study of Privatized Synthetic Data Generation Using Discrete Cosine Transforms

Kato Mivule
Computer Science Department
Bowie State University
Bowie Maryland, USA

***Abstract*—In order to comply with data confidentiality requirements, while meeting usability needs for researchers, entities are faced with the challenge of how to publish privatized data sets that preserve the statistical traits of the original data. One solution to this problem, is the generation of privatized synthetic data sets. However, during data privatization process, the usefulness of data, have a propensity to diminish even as privacy might be guaranteed. Furthermore, researchers have documented that finding an equilibrium between privacy and utility is intractable, often requiring trade-offs. Therefore, as a contribution, the Filtered Classification Error Gauge heuristic, is presented. The suggested heuristic is a data privacy and usability model that employs data privacy, signal processing, and machine learning techniques to generate privatized synthetic data sets with acceptable levels of usability. Preliminary results from this study show that it might be possible to generate privacy compliant synthetic data sets using a combination of data privacy, signal processing, and machine learning techniques, while preserving acceptable levels of data usability.**

***Keywords*—privatized synthetic data; Signal processing; Data privacy; discrete cosine transforms; Moving average filtering**

I. INTRODUCTION

Realizing an equilibrium between privacy and usability needs is a challenging undertaking that organizations have to engage in, to meet the terms of privacy regulations. To implement privacy acquiescent data transactions, trade-offs have to be made between privacy and usability requirements [1][2][3][4][5]. One way to address this problem, is the generation of privatized synthetic data sets that retain the statistical traits of the original data. Therefore, as a contribution, the Filtered Classification Error Gauge (Filtered x-CEG) methodology is suggested as a heuristic for the generation of privatized synthetic data [17]. The Filtered x-CEG is a variation of the Comparative x-CEG heuristic process described in Mivule and Turner (2013) [6] and [17]. The Filtered x-CEG heuristic works as follows: (i) Data privacy is applied to the data using noise addition; (ii) in the second step, signal processing technique of discrete cosine transforms, is used to mine the coefficients; (iii) the coefficients are added back to the noisy data; (iv) new privatized synthetic data is produced with a similar formation as the original[17]; (v) the moving average filter is then applied to the privatized synthetic data to improve usability; (vi) machine learning classification is used to test the filtered synthetic data for usability, with lower classification error (high classification accuracy) as an indication of better data usability [6][17]. Initial outcome from this study indicates that privatized synthetic data could be produced with adequate usability levels. Therefore, the main focus of this study is to employ data privacy, signal processing, and machine learning classification techniques in the generation of privatized synthetic data with acceptable levels of usability. The rest of the paper is organized as follows, in Section II, background and related work is given. Section III discusses the essential terms used in this paper, while Section IV focuses on the methodology. In Section V, the experiment is outlined and results discussion is done in Section VI. Finally in Section VII, the conclusion is given.

II. BACKGROUND AND RELATED WORK

In this section, a review of related work on using signal processing techniques for data privacy applications, is given [17]. While signal processing techniques have been applied for obfuscation in image and audio applications, there is not much work on using signal processing for specifically data privacy applications, such as, privatized synthetic data generation. However, of recent, researchers have picked up interest on applying signal processing techniques for data privacy implementations. For instance, on the use of signal processing in fulfilling data privacy challenges, Sankar, Trappe, Ramchandran, Poor, and Debbah (2013), noted that the necessary optimization task between data privacy and usability is a primary signal processing issue. Sankar et al., also observed there was a possibility of privacy assurances and solutions, by employing distributed signal processing methods [7]. Furthermore, Sankar et al., (2013), suggested the U-P trade-off region data privacy and utility signal processing based measurement model, for the quantification of data privacy and utility [7]. Consequently, usability, would be a measure of the closeness between the original and privatized data [7]. However, in this study, the classification error is used as a gauge for data privacy and usability quantification [6]. On the subject of discrete cosine transforms and data privacy, studies have mostly been done in the image and audio processing areas, with focus on access control instead of confidentiality [8][9][10][11]. In this paper, discrete cosine transforms methods are employed for data privacy applications, in this case, the generation of privatized synthetic data sets. Nevertheless, applications of Fourier transforms, for example discrete cosine transforms, were suggested by Mukherjee, Chen, and Gangopadhyay (2006) for the enhancement of privacy in Euclidean distance based clustering algorithms [11]. Mukherjee et al., (2006) observed that although original data allocations can be fittingly reconstructed in the confidential data, distance between points in the confidential data, is not conserved, thus clustering results with unsatisfactory performance [11]. At the same time Mukherjee et al., (2006) outlined advantages of employing Fourier transforms (discrete cosine transforms): (i) Conservation of Euclidean distance in the transformed data can be achieved, thus better clustering results; (ii) data compression could be attained by suppressing lesser coefficients and retaining greater coefficients; (ii) by suppressing coefficients, confidentiality of the data can be enhanced, thus making it complex for attackers to reconstruct the original data [11] [17]. In this study and in the suggested model, the suppression of coefficients as in Mukherjee et al (2006) model, is avoided. Rather, extraction of coefficients using discrete cosine transforms, and applying the coefficients in the generation of synthetic data with similar traits as the original, is done.

III. Essential Terms

While a number of data privacy and signal processing methods exist, it is beyond the span of this implementation paper to expansively survey each technique. The following are a description of some of the techniques used in this paper.

Noise addition: Random values are generated using the mean and standard deviation from the original data and added back to the original data, thus producing a confidential data set, using the following equation [12]:

$$\boldsymbol{Z = X + \varepsilon} \tag{1}$$

The symbol Z represents the confidential data, while X represents the original data, and ε symbolizes random values, chosen from a distribution of $\varepsilon \sim N(0, \sigma^2)$. The symbol ε represents an adjustable parameter, with a smaller ε producing data with traits much similar with the original, and a larger ε producing data that is much more dissimilar to the original [13]. In this paper, a normal distribution $\varepsilon \sim N(\mu = 1, \sigma = 0.2)$, is used to generate the noisy data that is then used in the signal processing, to generate coefficients which are then used to produce the privatized synthetic data set [17].

Discrete cosine transforms: Proposed by Ahmed, Natarajan, and Rao (1974), discrete cosine transform (DCT) is a process that converts a limited data sequence (real numbers) by summing up of cosine functions oscillating at different frequencies[13][14] [17]. DCT alters a set of real numbers $N: x_0, \dots, x_{n-1}$ into a set of real numbers $N: X_0, \dots, X_{n-1}$ using the following equation [14]:

$$\boldsymbol{X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right], k = 0, \dots, N-1.} \tag{2}$$

The symbol X_k, represents the set of altered data as a result of the DCT computation.

Moving Average Filter: In the moving average filter, each point in the output signal is a result of averaging a number of adjacent points in the input signal using the following formula: [16].

$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i+j] \tag{3}$$

The notation $x[i+j]$ symbolizes the input signal, while $y[i\ ]$ represents the output signal, and M stands for the number of points used in the moving average [16].

IV. Methodology

In this section, the methodology used in this paper and as described in [16], is outlined. The Filtered *x*-CEG, an adaptation of the Comparative *x*-CEG heuristic model outlined in Mivule and Turner (2013), is suggested [6]. Signal processing techniques, such as, discrete cosine transforms are used in the Filtered *x*-CEG, illustrated in Figure 1, unlike the model in [6], that does not involve signal processing methods [17]. The following are the steps involved in the generation of privatized synthetic data sets.

The Filtered *x*-CEG:

- *Step 1: Data privacy:* data privacy is implemented using noise addition – noisy data with statistical traits closer to the original is generated, with a normal distribution $\varepsilon \sim N(\mu = 1, \sigma = 0.2)$.
- *Step 2: Signal processing:* discrete cosine transforms is applied on the noisy data to extract coefficients.
- *Step 3: Synthetic data generation*: the obtained coefficients from Step 2, are added to the noisy data, producing a new confidential synthetic data set. The compensation from this phase is that it would be more difficult for an attacker to rebuild the original data; furthermore, the statistical traits from the original data could be preserved by using the acquired coefficients.
- *Step 4: Filtering*: The moving average filter is used in this phase, to reduce noise that could affect the usability of the data, with the aim for better data usability.
- *Step 5: Machine learning*: Machine learning is then applied on the privatized synthetic data to gauge for usability – with less classification error as an indicator of better data usability.
- *Step 6: The threshold*: if the classification error satisfies the desired threshold, then better usability is achieved and the privatized synthetic data could be published.

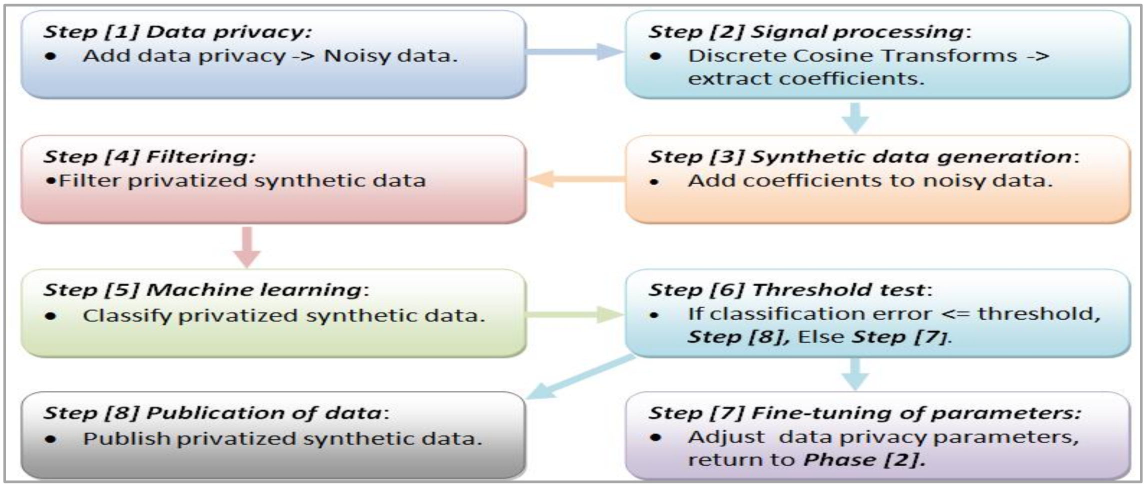

Step [1] Data privacy:
• Add data privacy -> Noisy data.
Step [2] Signal processing:
• Discrete Cosine Transforms -> extract coefficients.
Step [3] Synthetic data generation:
• Add coefficients to noisy data.
Step [4] Filtering:
•Filter privatized synthetic data
Step [5] Machine learning:
• Classify privatized synthetic data.
Step [6] Threshold test:
• If classification error <= threshold, Step [8], Else Step [7].
Step [7] Fine-tuning of parameters:
• Adjust data privacy parameters, return to Phase [2].
Step [8] Publication of data:
• Publish privatized synthetic data.

Fig. 1. The Filtered *x*-CEG process

- *Step 7: Fine tuning of parameters*: fine-tuning is done to the privacy parameters, and the signal processing is re-done if the threshold is not met. The procedure replicates x times until the preferred threshold is achieved, signifying improved data usability.
- *Step 8: Publication*: The privatized synthetic data with improved usability is published.

V. Experiment

The data used in this study comprised of the Fisher Iris data hosted at the UCI repository. The data contained 150 data items, four columns, the sepal length, sepal width, petal length, and petal width, with the fifth class column, representing the three classes, Setosa, Versicolar, and Virginica [15]. To produce the noisy data, the original data set was perturbed with noise addition at $N{\sim}(\mu = 1, \sigma = 0.2)$. This allocation of noise was selected since it mirrored statistical characteristics of the original data. After generation of the noisy data, discrete cosine transforms technique was used to obtain coefficients from the noisy data (which in this case was a close representation of the original data). The obtained coefficients were combined – added back to the noisy data, as illustrated in Figure 2, for an additional stratum of confidentiality, generating the privatized synthetic data set. The moving average filter was then used on the privatized synthetic data to remove excessive noise and thus increase usability. Machine learning classification was then applied on both the non-filtered and filtered privatized synthetic data. The following classifiers were used: Neural Networks, KNN, Naïve Bayes, Decision Trees, and AdaBoost Ensemble, employing a 10 fold cross-validation.

The threshold determination heuristic was then used by observing all classification errors and choosing data sets that met the threshold criteria. Only data sets that met the threshold criteria were published and statistical analysis performed on them.

VI. Results and Discussions

In this segment, outcome from the experiment on applying discrete cosine transforms (DCT) and filtering techniques for data privacy, is presented. Three groups of data results are observed: (i) original data, (ii) noisy data, and (iii) privatized synthetic data. A presentation of both descriptive and inference statistical results is also given.

A. Non-Filtered Privatized Sythentic DCT-based Data Results

Figures 3(a), 3(b), 3(c), and 3(d), represent results from the DCT process. In each graph of the illustrations, the lower data sequence represents the DCT coefficients, while the middle data sequence represents the noisy data, and the upper data sequence represents the generated privatized synthetic data. The DCT coefficients were mined from the noisy data and added to the same noisy data set, generating the privatized synthetic data. The noisy data was generated using very low noise addition to the original, to mimic the statistical properties of the original data. As can be seen in Figure 3(a), the privatized synthetic data sequence follows a similar pattern to the noisy data sequence, from an anecdotal view point. This could be an indication that it might be possible to generate privatized synthetic data sets that retain some statistical traits of the original data.

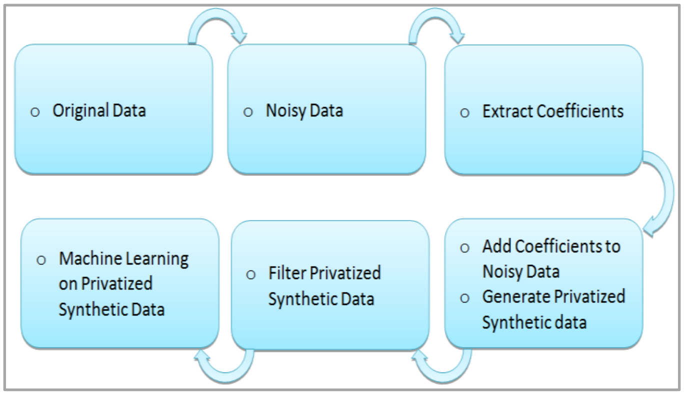

Original Data
Noisy Data
Extract Coefficients
Add Coefficients to Noisy Data
Generate Privatized Synthetic data
Filter Privatized Synthetic Data
Machine Learning on Privatized Synthetic Data

Fig. 2. Privatized synthetic data generation process

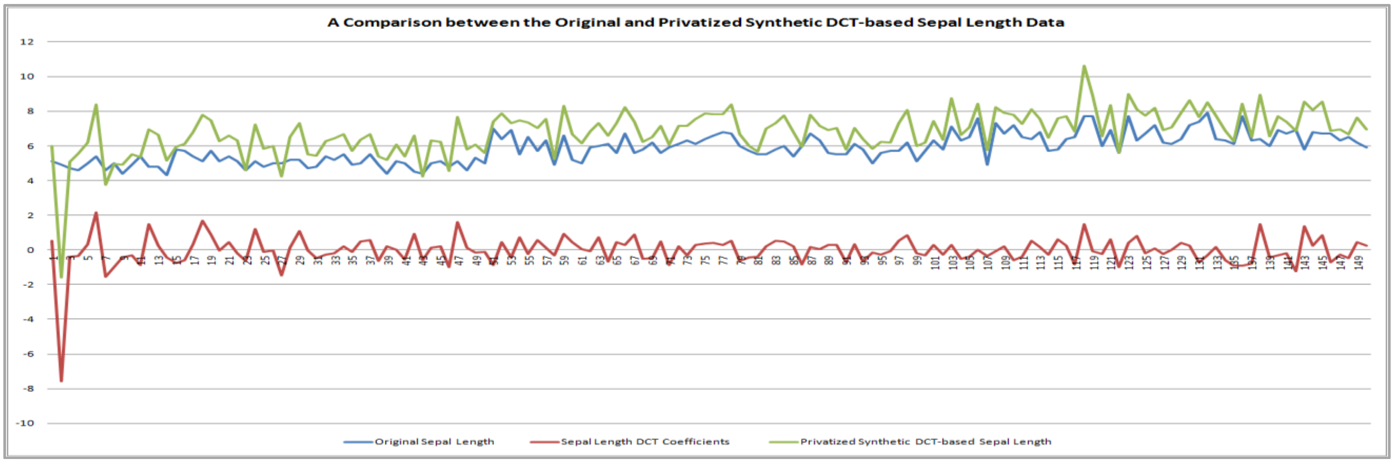


Fig. 3. (a)Privatized Synthetic Fisher-Iris data sequence – Sepal Length

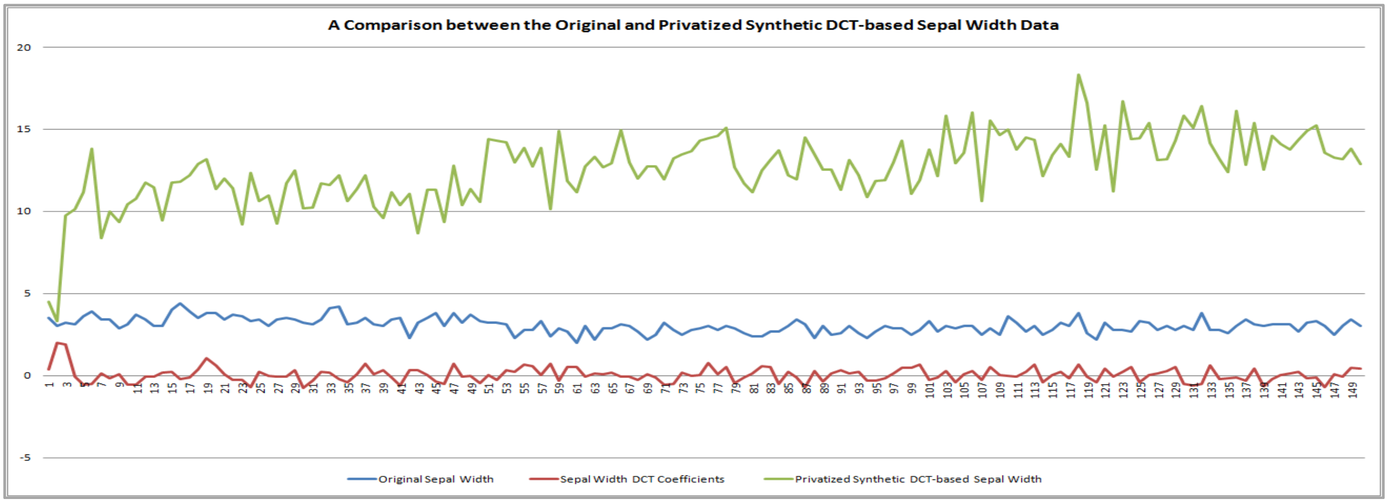


Fig. 3. (b)Privatized Synthetic data Fisher-Iris data sequence – Sepal Width

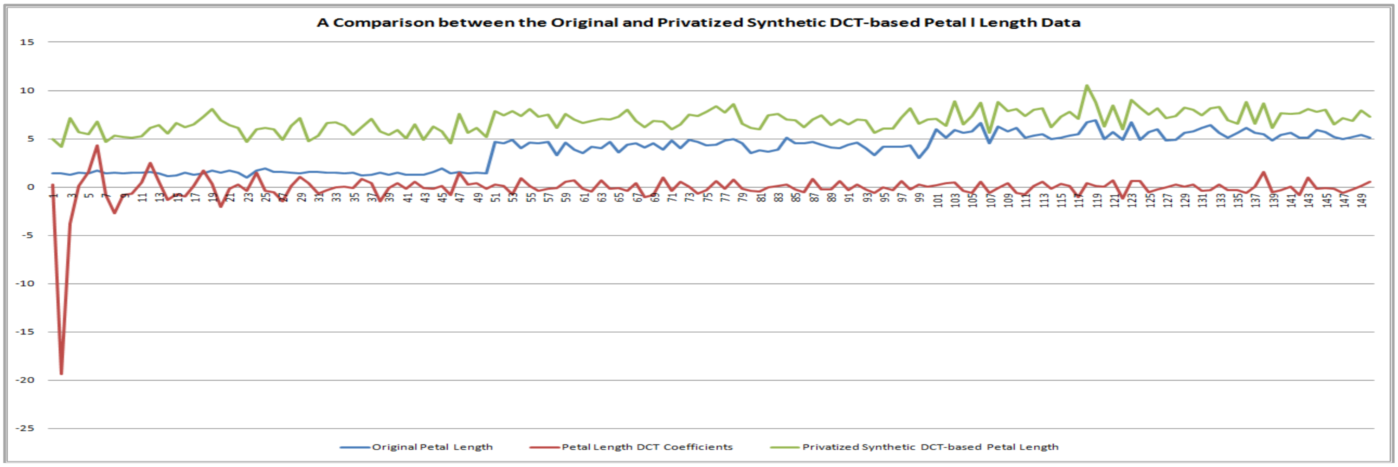


Fig. 3. (c)Privatized Synthetic data Fisher-Iris data sequence – Petal Length

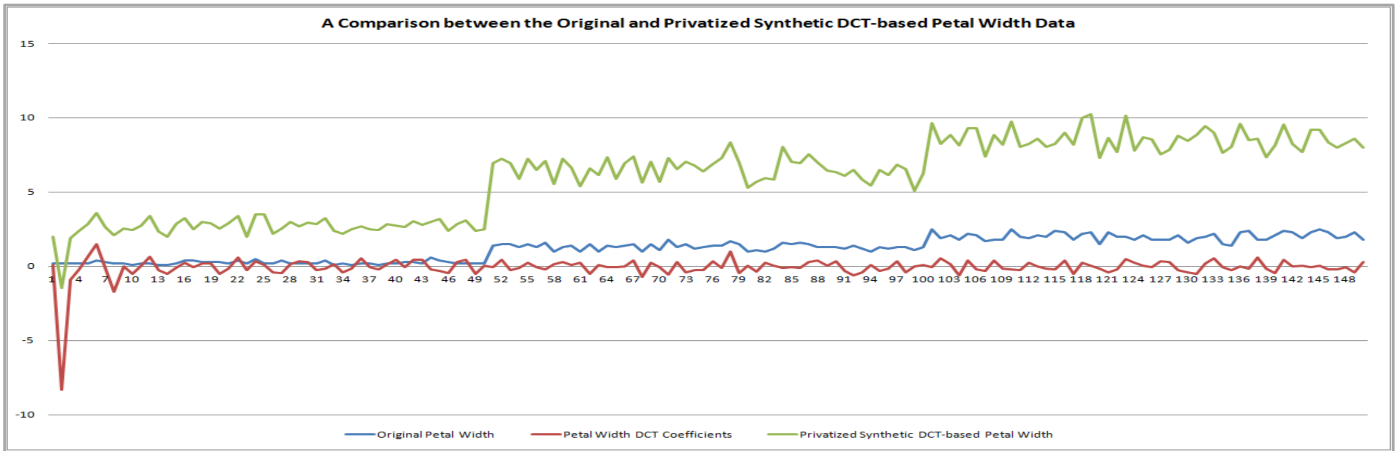


Fig. 3. (d)Privatized Synthetic data Fisher-Iris data sequence – Petal Width

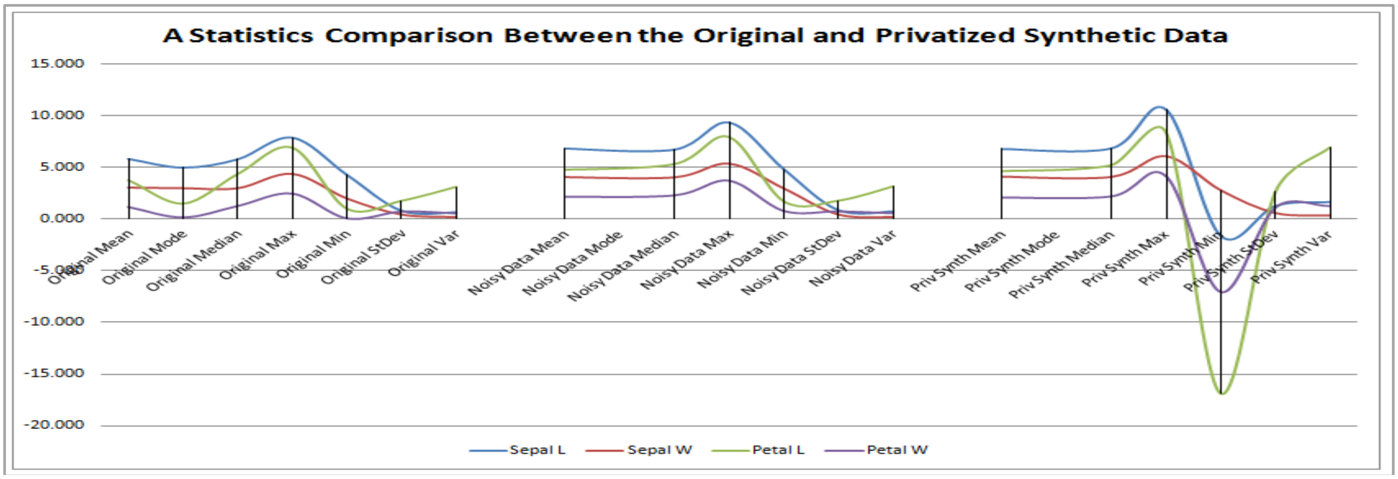


Fig. 4. Privatized Synthetic – descriptive statistics

As shown in Figures 3(b), 3(c), and 3(d), DCT-based privatized synthetic data did not automatically preserve the statistical skeletal structure of both the original and noisy data sets; and as further highlighted in Figure 4, with the descriptive statistics, a deformation of the original statistical skeletal structure occurred with the DCT-based privatized synthetic data. An anecdotal view of Figure 4 and Table I, show that the statistical skeletal structural likeness of the original data is kept in the noisy data.

TABLE I. NON-FILTERED PRIVATIZED SYNTHETIC DATA – DESCRIPTIVE STATISTICS

Statistics	Sepal L	Sepal W	Petal L	Petal W
Original Mean	5.843	3.054	3.759	1.199
Original Mode	5.000	3.000	1.500	0.200
Original Median	5.800	3.000	4.350	1.300
Original Max	7.900	4.400	6.900	2.500

Original Min	4.300	2.000	1.000	0.100
Original Stdev	0.828	0.434	1.764	0.763
Original Var	0.686	0.188	3.113	0.582
Noisy Data Mean	6.841	4.077	4.766	2.200
Noisy Data Mode	#N/A	#N/A	#N/A	#N/A
Noisy Data Median	6.744	4.060	5.323	2.333
Noisy Data Max	9.353	5.398	7.921	3.747
Noisy Data Min	4.846	2.978	1.716	0.819
Noisy Data Stdev	0.880	0.432	1.778	0.776
Noisy Data Var	0.775	0.186	3.162	0.603
Priv Synth Mean	6.801	4.124	4.632	2.125
Priv Synth Mode	#N/A	#N/A	#N/A	#N/A
Priv Synth Median	6.863	4.101	5.225	2.232
Priv Synth Max	10.608	6.115	8.356	4.173
Priv Synth Min	-1.603	2.799	-16.889	-7.010
Priv Synth Stdev	1.295	0.583	2.632	1.142
Priv Synth Var	1.677	0.340	6.926	1.305

However, the same statistical skeletal structure is deformed after applying DCT, in the privatized synthetic data. This could mean that simply adding noise addition to generate a noisy data set might not be enough, since an attacker could guess the original with a higher prospect of success. However, the statistical structure of the privatized synthetic data set is deformed when compared to the original and thus might make it more difficult for an attacker to guess the original composition while at the same time offering some usability to the end user of the privatized synthetic data set. Nevertheless, the mean of the privatized synthetic data is preserved when compared to the mean of the noisy data, as illustrated in Table I. For example, the mean of the noisy data is 6.841, whereas the mean of the privatized synthetic data is at 6.863 for the Sepal length class as recorded in Table I. Yet still, the median and max values are not preserved in the privatized synthetic data set. The covariance values between the noisy data and the privatized synthetic data sets are shown in Figure 5 and Table II. The standard deviation and covariance of the privatized synthetic data set is also not analogous to the noisy and original data. This might be good for privacy preservation in the privatized synthetic data set, while still maintaining some level of usability with the similar mean values.

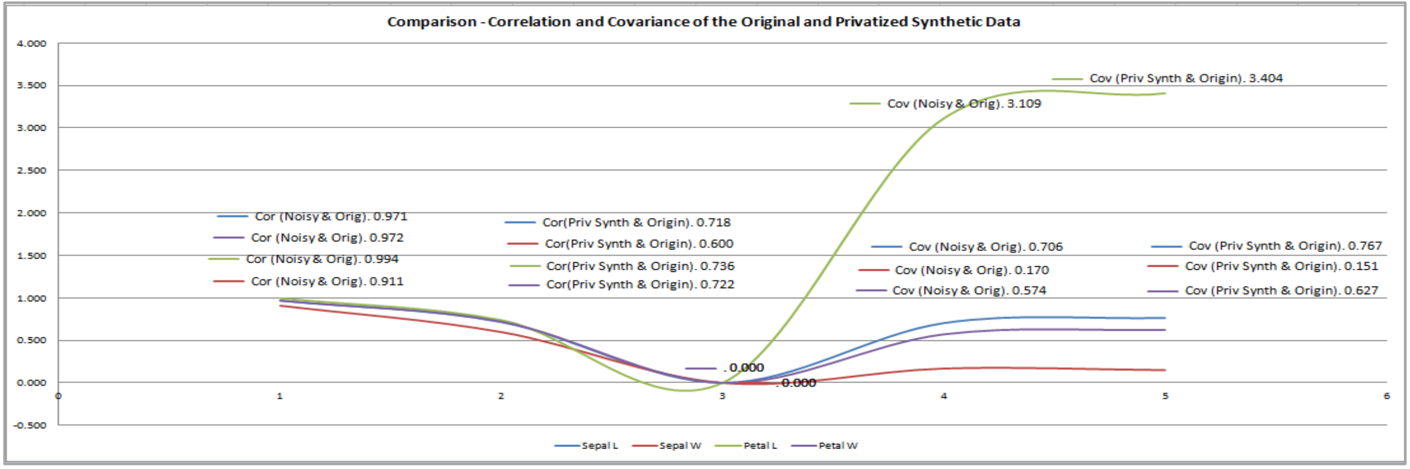


Fig. 5. Privatized Synthetic data – correlation and covariance

The results in Table II, show covariance values between 3.1 and 3.4, for the Petal length, and between 0 and 1, for the Sepal length, Sepal width, and Petal width, an indication of a diminutive inclination for the compared data to grow simultaneously.

TABLE II. NON-FILTERED PRIVATIZED SYNTHETIC DATA – CORRELATION AND COVARIANCE

Statistics	Sepal L	Sepal W	Petal L	Petal W
Correl (Noisy Data & Orig)	0.971	0.911	0.994	0.972
Correl (Synth & Orig)	0.718	0.600	0.736	0.722
Cov (Noisy Data & Orig)	0.706	0.170	3.109	0.574
Cov (Synth & Orig)	0.767	0.151	3.404	0.627

The correlation shown in Table II, between the noisy data and the original data, indicate results varying from 0.971 to 0.994, demonstrating a strong relationship. However, correlation results between the privatized synthetic and original data indicate a range of values from 0.060 to 0.74, signifying more or less a small relationship between the privatized synthetic data and the original data. Yet still, this could be good for privacy preservation even though a level of usability might be lost. Nonetheless, it might be said that DCT-based privatized synthetic data did not preserve the statistical traits of the original but did maintain the mean values. To investigate this premise further, DCT-based privatized synthetic data is passed through the filtering procedure.

B. *Filtered Privatized Sythentic DCT-based Data Results*

Results in Figures 6(a), 6(b), 6(c), and 6(d), represent the outcome of the experiment after applying filtering on the DCT-based privatized synthetic data. The lower sequence in each of the graphs shown in the illustrations, represents the DCT coefficients, while the middle sequence represents the noisy data, and the upper sequence represents the generated privatized synthetic data after applying filtering.

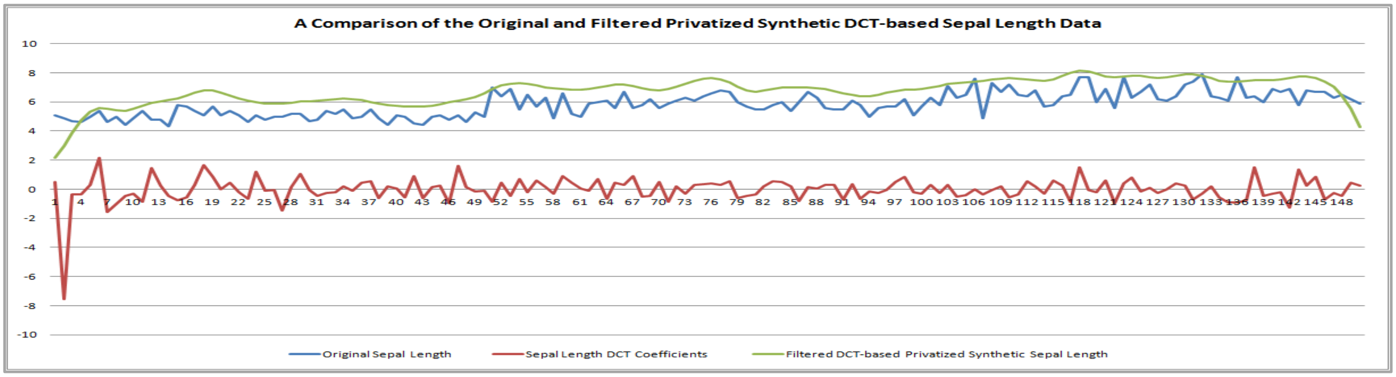


Fig. 6. (a)Filtered Privatized Synthetic Fisher-Iris data sequence – Sepal Length

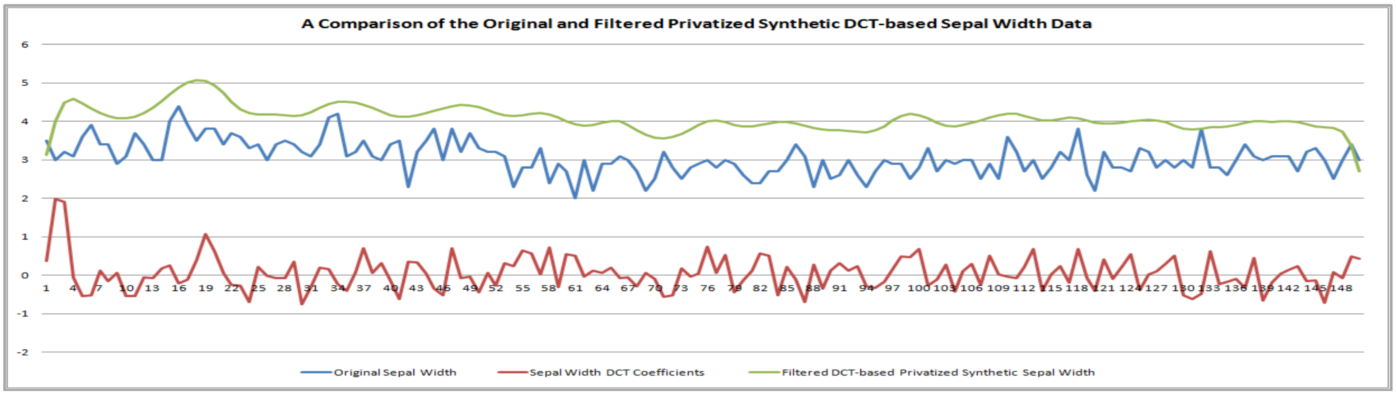


Fig. 6. (b)Filtered Privatized Synthetic Fisher-Iris data sequence – Sepal Width

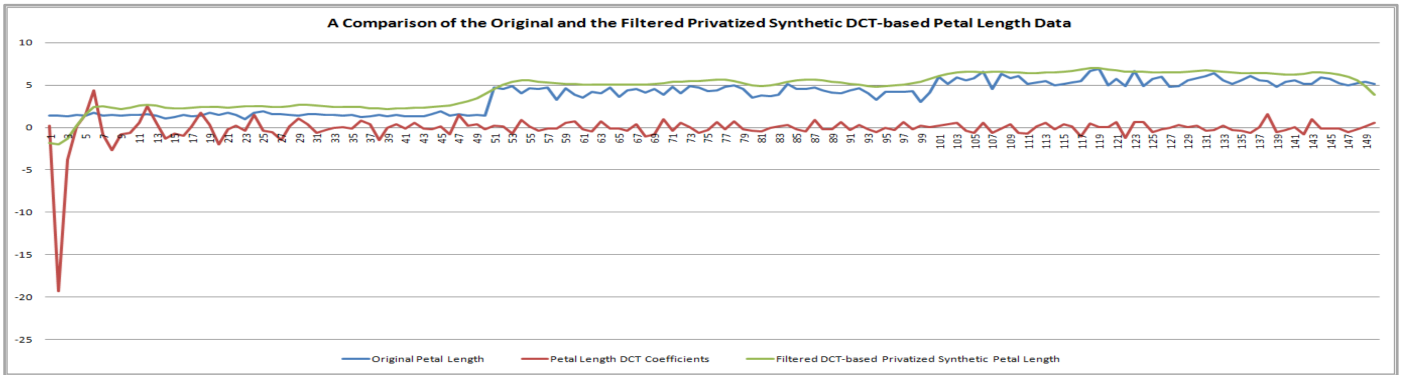


Fig. 6. (c)Filtered Privatized Synthetic Fisher-Iris data sequence – Petal Length

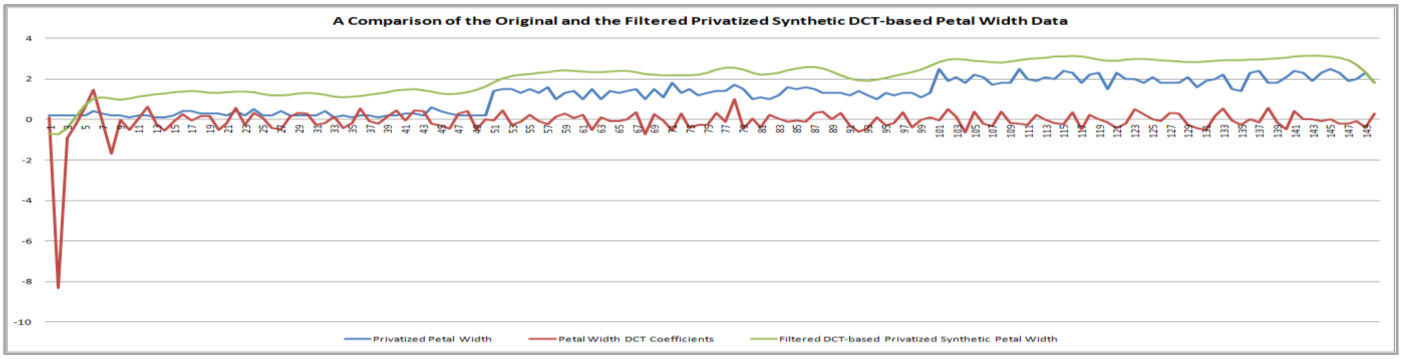


Fig. 6. (d)Filtered Privatized Synthetic Fisher-Iris data sequence – Petal Width

The moving average filtering with kernel width window of 4.0, was employed in the experiment. Regardless of the filtering process, Filtered privatized synthetic data did not preserve a good deal of the statistical traits and skeletal makeup of the noisy and original data, as illustrated in Figure 7; the results are similar to those produced for the non-filtered privatized synthetic data in Figure 4. However, the mean values were preserved in the Filtered privatized synthetic data, similar to results in the non-filtered privatized synthetic data, as shown in Table III. The outcome from this part of the study, indicates that although DCT based privatized synthetic data did not preserve some of the statistical traits, the mean values were maintained, an indication of some level of usability. Additionally, it might be possible that better privacy guarantees could be offered with DCT-based privatized synthetic data, and make it more challenging for an attacker to make precise deductions. Therefore, for the production of privatized synthetic data sets with less emphasis on data usability (utility), DCT-based privatized synthetic data sets might offer some interesting outcomes. However, there was a slight improvement in the correlation values, as shown in Figure 8. The filtered privatized synthetic data and the original data correlation values ranged from 0.5 to 0.9, compared to the 0.6 to 0.7 range of the non-filtered privatized synthetic data and the original.

TABLE III. FILTERED PRIVATIZED SYNTHETIC DATA – DESCRIPTIVE STATISTICS

Statistics	Sepal L	Sepal W	Petal L	Petal W
Original Mean	5.843	3.054	3.759	1.199
Original Mode	5.000	3.000	1.500	0.200
Original Median	5.800	3.000	4.350	1.300
Original Max	7.900	4.400	6.900	2.500
Original Min	4.300	2.000	1.000	0.100
Original StDev	0.828	0.434	1.764	0.763
Original Var	0.686	0.188	3.113	0.582
Noisy Data Mean	6.841	4.077	4.766	2.200
Noisy Data Mode	#N/A	#N/A	#N/A	#N/A
Noisy Data Median	6.744	4.060	5.323	2.333
Noisy Data Max	9.353	5.398	7.921	3.747
Noisy Data Min	4.846	2.978	1.716	0.819
Noisy Data StDev	0.880	0.432	1.778	0.776
Noisy Data Var	0.775	0.186	3.162	0.603
Priv Synthetic Mean	6.801	4.124	4.632	2.125
Priv Synthetic Mode	#N/A	#N/A	#N/A	#N/A
Priv Synthetic Median	6.863	4.101	5.225	2.232
Priv Synthetic Max	10.608	6.115	8.356	4.173
Priv Synthetic Min	-1.603	2.799	-16.889	-7.010
Priv Synthetic StDev	1.295	0.583	2.632	1.142
Priv Synthetic Var	1.677	0.340	6.926	1.305

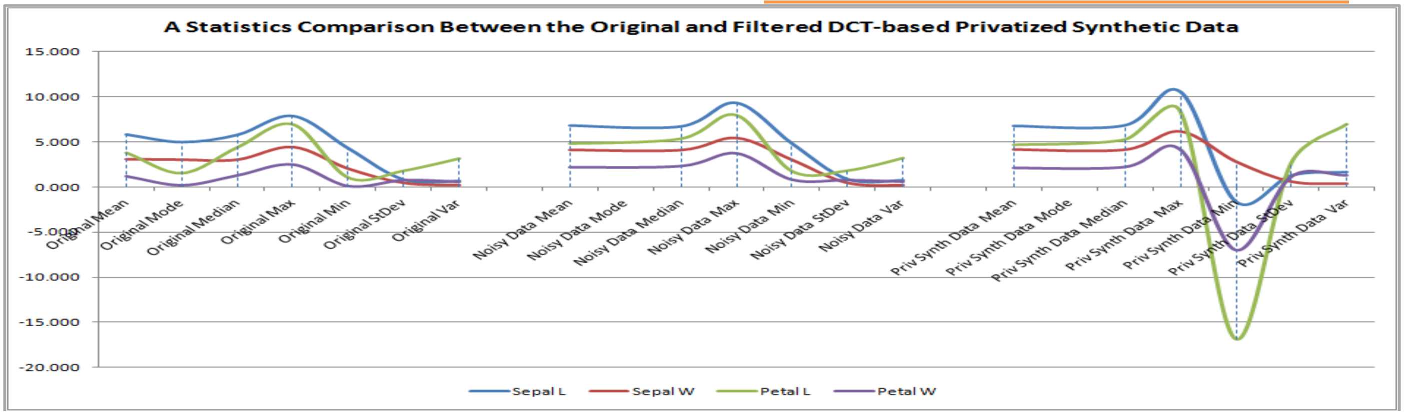


Fig. 7. Filtered Privatized Synthetic data descriptive statistics

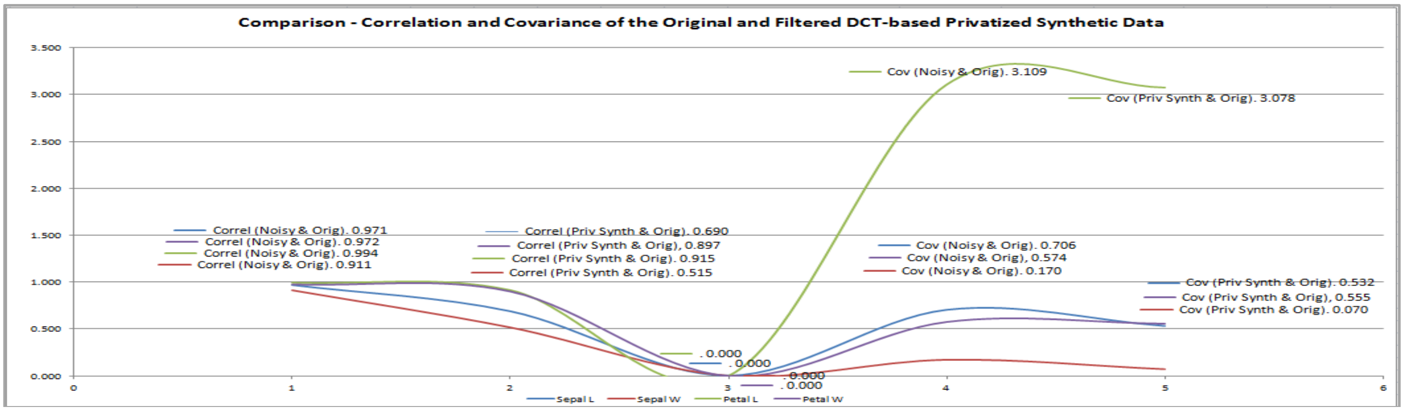


Fig. 8. Filtered Privatized Synthetic data – correlation and covariance

TABLE IV. FILTERED PRIVATIZED SYNTHETIC DATA – CORRELATION AND COVARIANCE

Statistics	Sepal L	Sepal W	Petal L	Petal W
Correl (Noisy Data & Orig)	0.971	0.911	0.994	0.972
Correlation(Priv Synth & Origin)	0.690	0.515	0.915	0.897
Cov (Noisy Data & Orig)	0.706	0.170	3.109	0.574
Cov (Priv Synth & Origin)	0.532	0.070	3.078	0.555

C. Machine Learning Classifier Results

Preliminary results from employing machine learning classification as a measure for data usability, are presented in this section. Both the non-filtered and filtered privatized synthetic data were sent through a chain of machine learning classifiers, namely, Neural Nets (NN), K-nearest Neighbor (KNN), Naïve Bayes (NB), Decision Trees (DT) – Random Forest, in this case, and AdaBoost ensemble. Each classifier returned the classification error, with a higher classification error signifying low data usability, and a low classification error representing improved data usability.

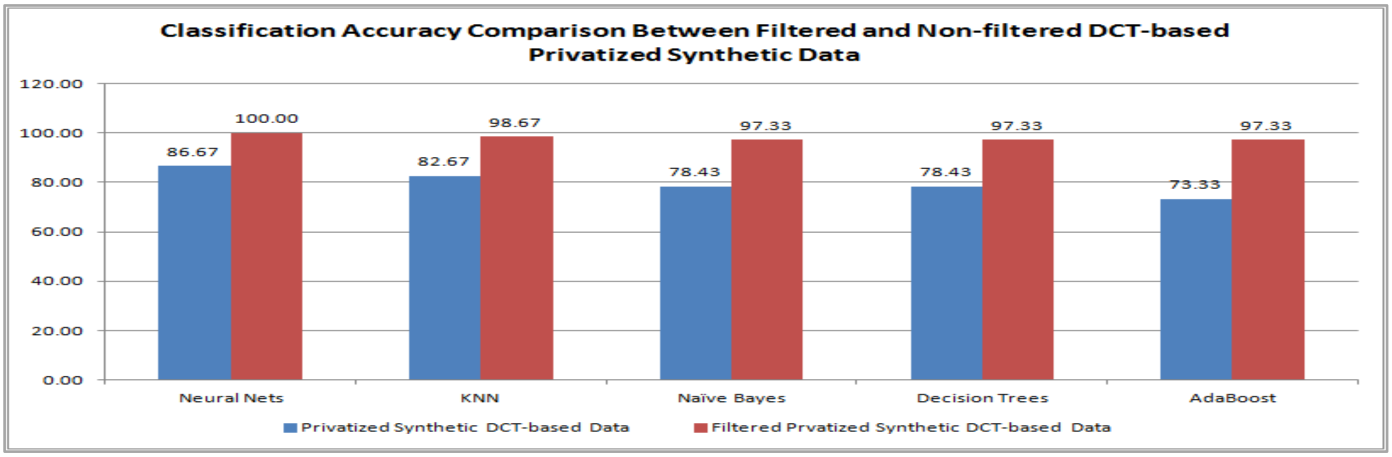


Fig. 9. Classification of Non-Filtered and Filtered data

In Figure 9 and Table V, classification accuracy results were reported – with high classification accuracy as an indication of low classification error and better data usability. However, low classification accuracy indicates higher classification error and likewise signifies low data usability.

TABLE V. CLASSIFICATION ACCURACY FOR BOTH NON-FILTERED AND FILTERED DATA

.Classifier	Privatized Synthetic DCT-based Data	Filtered Privatized Synthetic DCT-based Data
NN	86.67	100.00
KNN	82.67	98.67
NB	78.43	97.33
DT	78.43	97.33
AdaBoost	73.33	97.33

Experimental results, as indicated in Figure 9, Figure 10, and Table V, show that there was better performance with filtered privatized synthetic data, with returned higher classification accuracy results, and thus lower classification error. This signifies that filtering might have a profound effect on the classification accuracy of a perturbed data set. For instance, a look at the classification accuracy results, the non-filtered privatized synthetic data, returned a classification accuracy of 86.67 for NN, 82.67 for KNN, and 73.33 for AdaBoost. However, filtered privatized synthetic data returned 100.00 for NN, 98.67 for KNN, and 95.33 for AdaBoost, an indication that filtering does have an effect. The Neural Net classifier, represented by the top sequence in Figure 10, offered the best performance in terms of resilience, among classifiers used in this experiment, on both non-filtered and filtered privatized synthetic data. In general, there was a significant improvement in the performance of all classifiers after application of filtering as illustrated in Figure 10. Consequently, our preliminary results indicate that the technique of filtering noisy data might be significant in enhancing the classification accuracy of data, as such, improving data usability for privatized synthetic data sets. However, concerns about to what degree filtering has to be employed in privatized synthetic data generation, is still challenging. Secondly, inquiries about what quantity of information might be lost at some point in the filtering process, also remain legitimate.

D. Threshold Determination Results

Results in this section, as illustrated in Figure 11, show how the threshold was determined. To find out the threshold, a heuristic was employed by first, using the average value function to compute the mid-point values, and secondly, calculating the mean values [17]. As shown in Table VI, values used in the calculation of both the mid-point and mean were selected from the classification accuracy results. After selecting the mid-point and mean values, the threshold was then selected by taking the max value between the max mid-point and max mean values as shown in Table VI. From our preliminary results, the selected threshold value was 93.34 classification accuracy or 6.66 classification error. Any privatized synthetic data set that met this threshold requirement, was selected, as offering better data usability. Once the threshold is determined and privatized synthetic data set is chosen, the Filtered *x*-CEG procedure stops; the selected data sets that meet the threshold requirement are then published. Conversely, if the threshold criteria is not satisfied, and no data sets are chosen, then the Filtered *x*-CEG algorithm would proceed to the step of adjusting data privacy parameters and going through the classifier procedure again *x*-times, until the threshold criteria is satisfied.

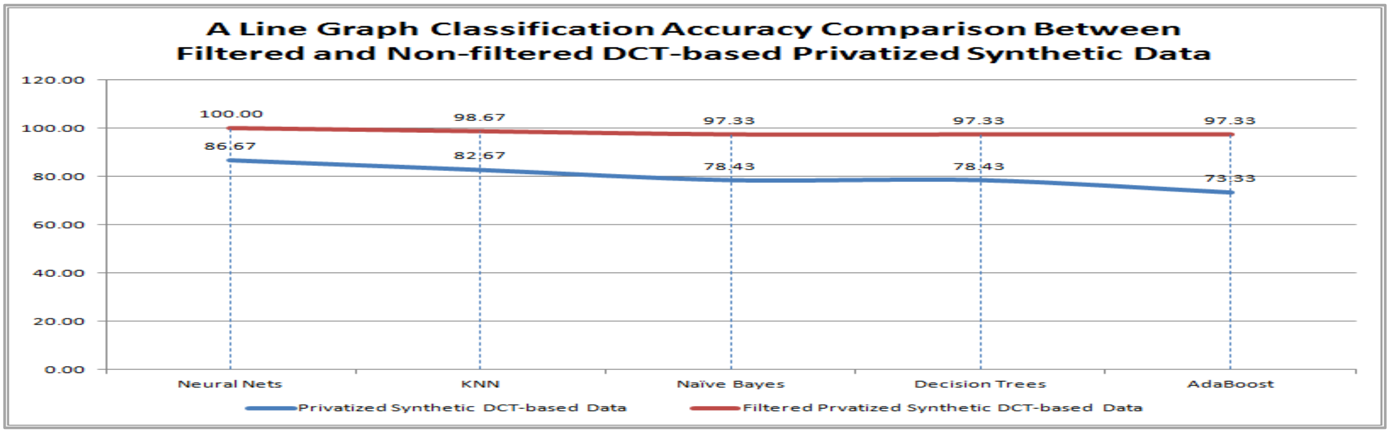
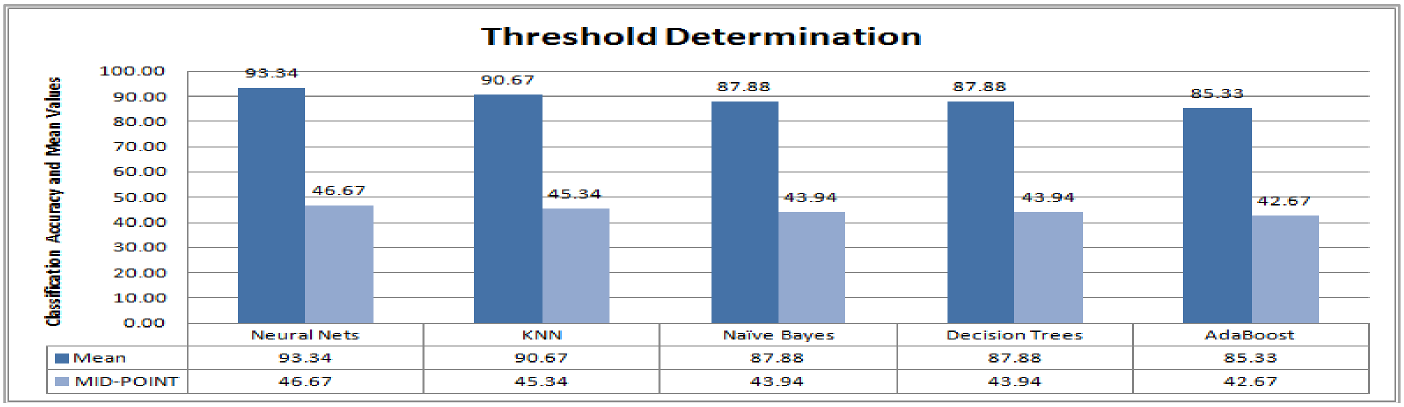


Fig. 10. Performance of classifiers on non-filtered and filtered data



	Neural Nets	KNN	Naïve Bayes	Decision Trees	AdaBoost
■ Mean	93.34	90.67	87.88	87.88	85.33
■ MID-POINT	46.67	45.34	43.94	43.94	42.67

Fig. 11. The mean and mid-point values

TABLE VI. DETERMINIG THE THRESHOLD

Priv Synth Data	NN	KNN	NB	DT	AdaBoost	*Max*
Mean	93.34	90.67	87.88	87.88	85.33	*93.34*
MID-POINT	46.67	45.34	43.94	43.94	42.67	*46.67*
Max	93.34	90.67	87.88	87.88	85.33	***93.34***

VII. CONCLUSION

In this investigation, the Filtered Classification Error Gauge (Filtered *x*-CEG) heuristic was presented and tested. The suggested data privacy model, in which data privacy, signal processing, and machine learning methods are employed to generate privatized synthetic data sets with satisfactory usability levels, was implemented. Preliminary outcome from this investigation indicates that signal processing techniques, such as, discrete cosine transforms, could be used in concert with data privacy techniques to produce privatized synthetic data sets in compliance with confidentiality requirements. Additionally, initial outcome from this study, indicates that filtering might have a corollary to the usability and performance of a privatized synthetic data set when classification is applied to the data set. Filtered privatized synthetic data returned higher classification accuracy results than the non-filtered privatized synthetic data, an indication that filtering might enhance usability of privatized data sets. On the other hand, non-filtered and filtered privatized synthetic data sets did preserve the mean but not the correlation with the original data, an indication of no relationship. In addition, non-filtered and filtered privatized synthetic data sets did not maintain the skeletal structure of the original data, a further indication of dissimilarity. Yet this dissimilarity might be beneficial for improved confidentiality, and perhaps signify that it might be possible to generate confidential synthetic data sets with enhanced usability, by maintaining some statistical traits of the original data, such as, the mean. The Moving Average Filtering procedure was employed in this investigation, using a kernel width window of size 4.0. While the Filtering might have an effect on improving the classification accuracy results, as we showed in the preliminary results, experimenting with various filtering methods not used in this investigation would be worthwhile. The question of what most effective signal processing procedure one would select for executing such a privatized synthetic data generating procedure, remains a case by case proposition and open to further investigation. Yet still, a variety of algorithms could be employed in the generation of confidential synthetic data with strong privacy guarantees, such as, differential privacy. Even more, finding the right equilibrium between privacy and usability requirements, remains challenging and any proposed solution would necessitate trade-offs on a case-by-case basis.

A. Limitations and Future Work

Because of the emergent challenge of big data, the extent and complexity of data confidentiality is at the same time, growing, and as such, it is outside the reach of this investigation to tackle each subject in the data confidentiality sphere. As such, the goal of this investigation was to look at privatized synthetic data generation, by employing data privacy, signal processing, and machine learning methods. The goal of this investigation was not focused on the type of attacks on the privatized synthetic data, a subject while important, is left for future work. The investigation was restricted to DCT transforms and the moving average filtering techniques. The Fisher-Iris data set was the only data set used in this study. Therefore, future works will comprise of testing generated privatized synthetic data against various adversary attacks, employing of various signal processing and filtering techniques, not used in this investigation, using other large data sets, finally application of various machine learning techniques not covered in this investigation.

ACKNOWLEDGMENT

Portions of this work were presented as part of the dissertation by the author in fulfillment of the requirements for the D.Sc. degree, in the Computer Science Department, at Bowie State University [17]. Special thanks and acknowledgement of Dr. Claude Turner and Dr. Soo-Yeon Ji, in the Computer Science Department, at Bowie State University, for the great and tireless efforts in making this work possible. Special thanks to the peer reviewers for their generous feedback. This study was facilitated and made possible by the HBGI Grant from the United States Department of Education.

REFERENCES

[1] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, “Minimality Attack in Privacy Preserving Data Publishing,” in In Proceedings of the 33rd international conference on Very large data bases (VLDB ’07), 2007, pp. 543–554.

[2] A. Meyerson and R. Williams, “On the complexity of optimal K-anonymity,” in Proceedings of the twentythird ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems PODS 04, 2004, pp. 223–228.

[3] H. Park and K. Shim, “Approximate algorithms for K-anonymity,” in Proceedings of the 2007 ACM SIGMOD international conference on Management of data SIGMOD 07, 2007, pp. 67–78.

[4] A. Krause and E. Horvitz, “A Utility-Theoretic Approach to Privacy in Online Services,” J. Artif. Intell. Res., vol. 39, pp. 633–662, 2010.

[5] Y. W. Y. Wang and X. W. X. Wu, Approximate inverse frequent itemset mining: privacy, complexity, and approximation. 2005.

[6] K. Mivule and C. Turner, “A Comparative Analysis of Data Privacy and Utility Parameter Adjustment, Using Machine Learning Classification as a Gauge,” Procedia Comput. Sci., vol. 20, pp. 414–419, 2013.

[7] L. Sankar, W. Trappe, K. Ramchandran, H. V. Poor, and M. Debbah, “The Role of Signal Processing in Meeting Privacy Challenges,” IEEE Signal Process. Mag., vol. 30, no. 5, pp. 95–106, 2013.

[8] M. Diephuis, S. Voloshynovskiy, O. Koval, and F. Beekhof, “DCT sign based robust privacy preserving image copy detection for cloud-based systems,” in 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012, pp. 1–6.

[9] N. V. Lalitha, G. Suresh, and P. Telagarapu, “Audio authentication using Arnold and Discrete Cosine Transform,” in 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012, pp. 530–532.

[10] M. Niimi, F. Masutani, and H. Noda, “Protection of privacy in JPEG files using reversible information hiding,” in 2012 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), 2012, no. Ispacs, pp. 441–446.

[11] S. Mukherjee, Z. Chen, and A. Gangopadhyay, “A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms,” VLDB J., vol. 15, no. 4, pp. 293–315, Aug. 2006.

[12] J. Kim, “A Method For Limiting Disclosure in Microdata Based Random Noise and Transformation,” in Proceedings of the Survey Research Methods, American Statistical Association,, 1986, vol. Jay Kim, A, no. 3, pp. 370–374.

[13] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” IEEE Trans. Comput., vol. 100, no. 1, pp. 90–93, 1974.

[14] G. Strang, "The discrete cosine transform." SIAM Review, vol 41, no. 1, pp. 135-147, 1999.

[15] K. Bache and M. Lichman, “Iris Fisher Dataset - UCI Machine Learning Repository.” University of California, School of Information and Computer Science., Irvine, CA, 2013.

[16] K. Mivule and C. Turner, "Applying Moving Average Filtering for Non-interactive Differential Privacy Settings", Procedia Computer Science, (In Print), 2014, Philadephia, PA, USA

[17] K. Mivule, "An Investigation of Data Privacy and Utility Using Machine Learning as a Gauge", D.Sc. Dissertation, Computer Science Dept., Bowie State University. 2014: 262 pages; ProQuest: 3619387.

Vector Autoregression (Var) Model for Rainfall Forecast and Isohyet Mapping in Semarang – Central Java – Indonesia

Adi Nugroho
Faculty of Information Technology
Satya Wacana Christian University
Salatiga, Indonesia

Sri Hartati
Faculty of Mathematics and Natural Science
Gadjah Mada University
Jogyakarta, Indonesia

Subanar
Faculty of Mathematics and Natural Science
Gadjah Mada University
Jogyakarta, Indonesia

Khabib Mustofa
Faculty of Mathematics and Natural Science
Gadjah Mada University
Jogyakarta, Indonesia

***Abstract*—Agricultural and plantation activities in Indonesia, especially in Semarang, Central Java, Indonesia rely on water supply from the rainfall. The rainfall in the future is basically influenced by rainfall patterns, humidity and temperature in the past. In this case, Vector Autoregression (VAR) multivariate model is applied to forecast the rainfall in the future, in which all along Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) generally uses ARIMA model (*Autoregressive Integrated Moving Average*) to carry out the same thing. The study applied the data, comprising the data of rainfall, humidity and temperature taken on a monthly basis during 2001-2013 periods from 5 measurement stations. Plotting of rainfall forecast result with VAR method is portrayed in the form of isohyet contour map to see the correlation between rainfall and coordinates of the area of the rainfall. The forecast result shows that VAR method is quite accurate to use for rainfall forecast in the study area as well as better than ARIMA method to forecast the same thing as having smaller Mean Absolute Error (MAE) and Mean Absolute Percentage Error(MAPE).**

Keywords*—*Rainfall Forecast; VAR; Multivariate Time Series; Isohyet

I. Introduction

Indonesia has abundant natural resources in tropical areas with quite high rainfall as the islands are surrounded by the vast oceans, fairly high daily temperature and humidity [20]. Currently, there are approximately 40.6 million hectares of agricultural and plantation areas in Indonesia [22] which mostly rely the water supply on the rainfall. Regarding this matter, the western and northeast parts of Indonesia have geological condition and fertile soil which enables the agriculture/plantation can virtually be done as long as the water supply from the rainfall is sufficient [21].

Semarang, Central Java, Indonesia (the study area) is geographically located in Java island in the western part of Indonesia. The study area lies on the geographic position of 6º 5' – 7º 10' S and 110º 34' – 110º35' E with a total area of 37.366.838 hectares or about 373.7 km^2[19]. Generally, the rainfall in the study area follows the pattern of 2 seasons, namely dry season (April – September) and wet season (October – March)

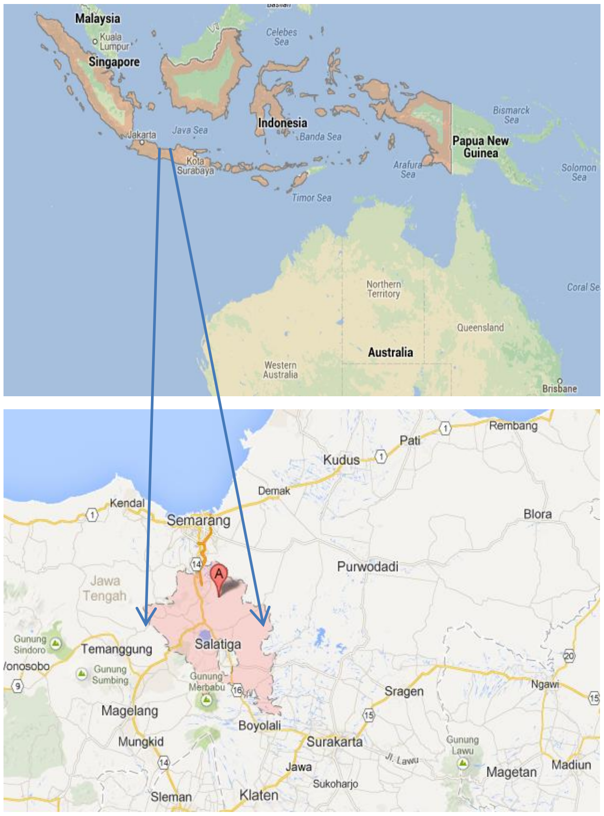


Fig. 1. Study Area Map

[20]. The conducted study tried the seasonal rainfall forecast (dry and wet season) 1 year forward (2014) based on monthly rainfall data taken along the span of the previous 13

years (2001-2013).The rainfall forecast is conducted by using multivariate time series method, or more particularly: **Vector Autoregression (VAR).** Furthermore, for the sake of decision making accuracy, the rainfall forecast is portrayed in the form of isohyet contour map to see the portrait of rainfall amounts in each sub-district in Semarang.

II. Previous Study

VAR method that is introduced by **Christopher A. Sims,** a Nobel Prize winner in econometrics, is previously pretty much used to develop econometric models [5] [13], such as to see the correlation of Gross Domestic Product (GDP) to the inflation level in a country, the tourist arrival level to a country which is influenced by many factors, the correlation between Composite Stock Price Index (Indeks Harga Saham Gabungan/IHSG), currency exchange rates and prices [11], the market response to the marketing mix [16], the correlation between public expenditure and economic growth [1], etc. For natural phenomena, especially for rainfall forecasting, Dewi Retno, et al [14] have conducted a study of correlation between rainfall in a region and rainfall in other nearby regions. Meanwhile, in the study that we conducted, the rainfall is connected to the rainfall in the previous periods as well as its correlation with humidity and temperature data.

In general, rainfall (precipitation) is a part of hydrologic cycle (water cycle) consisting of: (1) evaporation and/or evapotranspiration, (2) precipitation, and (3) surface water flow [12]. At each stage, the air humidity (water vapor percentage in an air volume)and the temperature would be very influential. When the temperature is relatively high due to the sunlight, there would be an evaporation/evapotranspiration of surface water/vegetation and water vapor in an air volume would be formed in which this water vapor in certain altitude, in turn, would form a core of condensation to form clouds. Due to the influence of low temperature and relatively immense droplet, the clouds then would drop back to the ground in the form of rain, snow, dew, fog, etc., which in turn would form surface water flowto restart the hydrologic cycle.

In Indonesia, the studies related to the rainfall according to Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) were mostly conducted using ARIMA method in which the study area has 22,13% MAPE value [6]. The study that we conducted aims to find out if VAR method that we applied can reduce this MAPE value, where the result would enhance the prediction accuracy.Moreover, in our study, we intend to connect the values of rainfall forecast to geographic coordinates of the area where the rainfall would drop by portraying it in the form of isohyet contour map.

III. Research Methodology

Time series is basically a measurement data taken in chronological order in certain time [9]. In the conducted study, based on the characteristic of each time series with some different kinds of data (rainfall, humidity and temperature), Vector Autoregression (VAR) method is applied. VAR is basically a combination of Autoregressive (AR) method and frequently known as Box-Jenkins method as developed by George Box and Gwilym Jenkins in 1976 [9].

For instance, the following is the time series of AR.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_{p-1} y_{t-p+1} + \beta_p y_{t-p} + \varepsilon \varepsilon_t \qquad (1)$$

In which y_t is the current value, while p is *lag* in *autoregressive* process; y_{t-1} to y_{t-p} is the measurement values from t-1 to t-p; β_0 is *intercept* value and $\beta_1$to β_p is regression coefficient from t-1 to t-p; and $\varepsilon\varepsilon_t$ is *error* value or frequently known as *white noise* which is assumed to be a normal distribution, independent of y_{t-1} and constant variance of σ^2 or equal to 0 [3] [7]. In terms of the use of VAR model in this conducted study, it requiredapplying stationary condition criteria, which is defined as a condition where its mean and variance are constant and the covariance is not time-dependent [11]. The stationarity in this study was tested using unit roottest with Augmented Dickey-Fuller (ADF) method. The ADF test has the following mathematical equation [8].

$$\Delta y_t = \beta_0 + \theta\, y_{t-1} + \sum_{i=2}^{p} \Phi_i \, \Delta y_{t-1+i} + \varepsilon \varepsilon_t \qquad (2)$$

In which Δy_t is time series value at the-t time minus time series value in 1 previous measurement period (the-t-1 time), θ is constant-valued $(\beta_1 + \cdots + \beta_{p-1} - 1)$[8] which is used to determine whether or not the unit roots exist with hypothesis $H_0 : \theta = 0$ (the data contain unit roots) and $H_1 : \theta < 0$ (the data do not contain unit roots). Meanwhile, Φ_i is trend coefficient on the time series data of which the value is equal to$\Phi_i = -\sum_{j=1}^{p} \beta_j$ [8]. In this case, the criteria that can be drawn are non-stationary data will have unit roots, while stationary data will not have unit roots. If the data are not stationary, it required to conduct differencing process once or several times on the related data until the data turn out to be stationary [10] [17].

Prior to VAR model was completely formed eventually, the accuracy level should be evaluated by calculating its lag value, which is generally indicated by its p-value. In the conducted study, in order to assess the feasibility level of the rainfall forecast model, it required applying Aikake's Information Criterion (AIC) calculation for some k independent variables where the AIC value is generally defined using the following mathematical equation [15].

$$AIC = log\sigma_k^2 + \frac{n+2k}{n} \ldots \qquad (3)$$

Where$\sigma_k^2 = \frac{SSE}{n}$ with $SSE = \sum_{i=1}^{n} (y_i - y_r)^2$

In which y_i is observed value at the-i time; k is the number of parameters in the model; $y_r$is *mean*; and n is the number of observation times. In this case, it can be stated that in case the AIC calculation value is smaller, the taken *lag* value is the better lag value [5] [15] as well as can be used as forecasting basis.

After accomplishing to determine the (p) lag value with AIC, the VAR mathematical equation system which theoretically does not distinguish the number of dependent and independent variables [17], the combination of rainfall, humidity and temperature variables by considering the accuracy of autoregressive equation (equation (1)) can be noted in the form of a matrix equation as follows [5] [13].

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} a_{1,1}^1 & a_{1,2}^1 & a_{1,3}^1 \\ a_{2,1}^1 & a_{2,2}^1 & a_{2,3}^1 \\ a_{3,1}^1 & a_{3,2}^1 & a_{3,3}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_{1,1}^p & a_{1,2}^p & a_{1,3}^p \\ a_{2,1}^p & a_{2,2}^p & a_{2,3}^p \\ a_{3,1}^p & a_{3,2}^p & a_{3,3}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ y_{3,t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon\varepsilon_{1,t} \\ \varepsilon\varepsilon_{2,t} \\ \varepsilon\varepsilon_{3,t} \end{bmatrix} \quad \ldots(4)$$

The matrix equation above (equation (4)), using regression notation can also be noted as follows (for rainfall).

$$y_{1,t} = c_1 + a_{1,1}^1 y_{1,t-1} + a_{1,2}^1 y_{2,t-1} + a_{1,3}^1 y_{3,t-1} + \ldots + a_{1,1}^p y_{1,t-p} + a_{1,2}^p y_{2,t-p} + a_{1,3}^p y_{3,t-p} + \varepsilon\varepsilon_{1,t} \quad (5)$$

In which $y_{1,t}$ is the rainfall at the-t time; $y_{2,t}$ is the humidity at the-t time; and $y_{3,t}$ is the temperature at the-t time. Meanwhile, c is the constant indicating the intercept; ε is the errors level; and p is the lag length. In this case, the parameter values $a_{1,1}^1$ to $a_{3,3}^p$ can be estimated by using Ordinary Least Square (OLS) method, by minimizing the value of squared error (minimizing ε^2 value) [13] [17]. The determination of the parameter values can be started from the-p matrix, and then recursively defined to the other parameters.

Once the best model was obtained and able to be used for forecasting, the forecast accuracy level of the model can also be evaluated mathematically using the following mathematical equation [5] [15] [17].

- Computing Mean Absolute Error (MAE)

MAE is a computation of mean absolute error to see how close the values between the forecast and the real value. MAE is generally defined as the following equation.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|F_t - Y_t| \quad (6)$$

Where F_t is the forecast value; Y_t is the actual data; and n is the length of time series of observation.

- Computing Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - F_t}{Y_t}\right| 10 \quad (7)$$

Where F_t is the forecast value and Y_t is the actual data.

In terns of the MAE and MAPE computation, the good model will have the smallest possible value of MAE as well as MAPE (less than or equal to 10%) [5] [15].

After the rainfall forecast value for the following year was accomplished, in order to portray the rainfalls in the study area related to the geographical position which is different from the measurement station, the isohyet map of Semarang then needs to be composed. The isohyet map is actually a regular contour map drawn to connect the dots in Semarang map with the same rainfall (forecast) values [18].

IV. Results And Discussion

The rainfall, humidity and temperature data used in this study were taken from Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) of Central Java, Indonesia. The obtained data are the data of monthly rainfall throughout 2001-2013 periods from 5 measurement stations, comprising Ungaran-Semarang, Semarang Kota, Bringin-Salatiga, Adisumarmo-Boyolali, and Borobudur-Magelang. At each measurement station, the rainfall forecast was carried out using VAR model, then (after the data from those 5 measurement stations were processed) the isohyet map for dry and wet season was composed by computing coordinates/geographic location of each station.

In order to apply VAR method to the existing data, we necessarily figure out the stationarity of data in each station initially using ADF calculation as the VAR[15]. ADF calculation for each time series data and interpretation of stationarity in each station are shown in Table 1.

TABLE I. Rainfall, Humidity and Temperature Stationarity Test in Various Measurement Stations

Rainfall Station	ADF Of Rainfall	Hypothesis	ADF Of Humidity	Hypothesis	ADF of Temperature	Hypothesis
Ungaran Semarang	-8.8005	S	-4.6565	S	-5.8516	S
Bringin-Salatiga	-3.0617	S	-3.9083	S	-3.9435	S
Adisumarmo Boyolali	-2.8473	S	-3.7498	S	-3.4986	S
Borobudur Magelang	-3.6135	S	-3.7498	S	-4.1076	S
Semarang Kota	-3.0887	S	-5.0996	S	-3.4963	S

Note: S (stationary hypothesis) and not requiring differencing process.

TABLE II. Table of AIC for Various VAR Computations

	VAR (3)	VAR (4)	VAR (5)	VAR (6)	VAR (7)	VAR (8)
Ungaran - Semarang	35	32	30	27	34	37
Bringin - Salatiga	25	23	20	17	22	26
Adisumarmo - Boyolali	26	24	21	17	23	27
Borobudur - Magelang	29	25	23	20	23	26
Semarang Kota	28	24	22	21	23	25

By following the stages that have been described previously, based on the rounded AIC calculation (Table 2), the final VAR model which can be obtained is VAR (6). VAR (6) with the smallest AIC value (shaded area) can be represented using mathematical equation model obtained through Ordinary Least Square (OLS) approach in the following equation (4).

- For measurement station of Ungaran-Semarang, $Y_{1,t}$ = -140,785646 + 0.182222205 $Y_{1,t-1}$ - 0.01409596 $Y_{1,t-2}$ - 0.09658367 $Y_{1,t-3}$ - 0.100907471 $Y_{1,t-4}$ - 0.118368817 $Y_{1,t-5}$ - 0.118000286 $Y_{1,t-6}$ + 0.019826927 $Y_{2,t-1}$ + 0.2879448 $Y_{2,t-2}$ - 0.04463891 $Y_{2,t-3}$ - 0.013124160 $Y_{2,t-4}$ - 0.017253552 $Y_{2,t-5}$ + 0.139557736 $Y_{2,t-6}$+ 0.01099009 $Y_{3,t-1}$ + 0.1034915 $Y_{3,t-2}$ + 0.48754994 $Y_{3,t-3}$ - 0.104888805 $Y_{3,t-4}$ + 0.032615013 $Y_{3,t-5}$+ 0.131957736 $Y_{3,t-6}$.
- For measurement station of Bringin – Salatiga, $Y_{1,t}$ = -349,0890 + 0.29781892 $Y_{1,t-1}$ - 0.56327709 $Y_{1,t-2}$ + 0.03872342 $Y_{1,t-3}$ - 0.111907471 $Y_{1,t-4}$ -0.107368817 $Y_{1,t-5}$ - 0.107000286 $Y_{1,t-6}$ - 0.08397126 $Y_{2,t-1}$ + 0.29326751 $Y_{2,t-2}$ - 0.25521153 $Y_{2,t-3}$ - 0.012024160 $Y_{2,t-4}$ - 0.017253552 $Y_{2,t-5}$ + 0.139557736 $Y_{2,t6}$+ 0.29648620 $Y_{3,t-1}$ + 0.01739839 $Y_{3,t-2}$ - 0.10278834 $Y_{3,t-3}$ - 0.103788805 $Y_{3,t-4}$ + 0.031515013 $Y_{3,t-5}$ + 0.139557736 $Y_{3,t-6}$.
- For measurement station of Adi Sumarmo - Boyolali, $Y_{1,t}$ = -468,738158 - 0.490667870 $Y_{1,t-1}$ - 0.095312517 $Y_{1,t-2}$ - 0.033233273 $Y_{1,t-3}$ - 0.122907471 $Y_{1,t-4}$ - 0.125368817 $Y_{1,t--5}$ - 0.129000286 $Y_{1,t-6}$ + 0.42273818 $Y_{2,t-1}$ - 1.18946346 $Y_{2,t-2}$ - 0.47307387 $Y_{2,t-3}$ - 0.012244160 $Y_{2,t-4}$ - 0.017475552 $Y_{2,t-5}$ + 0.151757736 $Y_{2,t-6}$ + 3.2756840 $Y_{3,t-1}$ + 1.8152582 $Y_{3,t-2}$ + 2.2411575 $Y_{3,t-3}$ - 0.103999805 $Y_{3,t-4}$ + 0.053715013 $Y_{3,t-5}$ + 0.158579936 $Y_{3,t-6}$.
- For measurement station of Borobudur - Magelang, $Y_{1,t}$= - 272,4933 + 0.04610669 $Y_{1,t-1}$ - 0.43096825 $Y_{1,t-2}$ - 0.35201372 $Y_{1,t-3}$ - 0.100907471 $Y_{1,t-4}$ - 0.107368817 $Y_{1,t-5}$ - 0.107000286 $Y_{1,t-6}$ + 0.99235069 $Y_{2,t-1}$ - 0.52613470 $Y_{2,t-2}$ - 1.16998137 $Y_{2,t-3}$ - 0.018084160 $Y_{2,t-4}$ - 0.015033552 $Y_{2,t-5}$ + 0.139225736 $Y_{2,t6}$ - 0.18096627 $Y_{3,t-1}$ - 0.22172650 $Y_{3,t-2}$ + 0.17670065 $Y_{3,t-3}$ - 0.103566805 $Y_{3,t-4}$ + 0.031515123 $Y_{3,t-5}$ + 0.139535536 $Y_{3,t-6}$.
- For measurement station of Semarang Kota, $Y_{1,t}$ = -638,137877 + 0.51796624 $Y_{1,t-1}$ - 0.52963239 $Y_{1,t-2}$ + 0.08272823 $Y_{1,t-3}$ - 0.100907471 $Y_{1,t-4}$ -0.107368817 $Y_{1,t-5}$ - 0.107000286 $Y_{1,t-6}$ - 0.10935264 $Y_{2,t-1}$ + 0.28539208 $Y_{2,t-2}$ - 0.13033554 $Y_{2,t-3}$ -- 0.046224160 $Y_{2,t-4}$ - 0.011693552 $Y_{2,t-5}$ + 0.137787736 $Y_{2,t-6}$ + 0.76188310 $Y_{3,t-1}$ + 1.93032475 $Y_{3,t-2}$ - 1.25373571 $Y_{3,t-3}$ - 0.103722205 $Y_{3,t-4}$ + 0.031559413 $Y_{3,t-5}$ + 0.139591136 $Y_{3,t-6}$.

Where

- $Y_{1,t}$ is rainfall value at the-t time.
- $Y_{2,t}$ is humidity value at the-t time.
- $Y_{3,t}$ is temperature value at the-t time.

TABLE III. MAE AND MAPE CALCULATION

Rainfall Station	MAE	MAPE (%)
Ungaran - Semarang	6.95128	2.139876
Bringin - Salatiga	13.15702	6.008978
Adisumarmo - Boyolali	10.23671	6.558561
Borobudur - Magelang	13.02692	5.713805
Semarang Kota	12.36196	6.707421

For calculating the accuracy of data processing, the calculation of MAE and MAPE value for each measurement station of which the result can be seen in Table 3 above is required. In general, the data processing result present the quite well values (the relatively small value of MAE and MAPE value that are below the range of 10%), hence VAR (6) model can be stated to be pretty well to forecast the rainfall in the study area (Semarang).

TABLE IV. RAINFALL FORECAST USING VAR (6) METHOD

Rainfall Station	Geographic Coordinate	Rainfall Forecast of Dry Season (mm/6 month)	Rainfall Forecast of Wet Season (mm/6 month)
Ungaran Semarang	426024.61, 9206491.87	1001.3251	2254.8765
Bringin Salatiga	448470.46, 9201746.57	1128.3067	2000.9843
Adisumarmo Boyolali	473139.18, 9168238.21	1171.3422	2222.9706
Magelang	412328.38, 9158961.91	1171.3422	2472.0129
Semarang Kota	435567.69, 9227421.87	994.2415	2230.9723

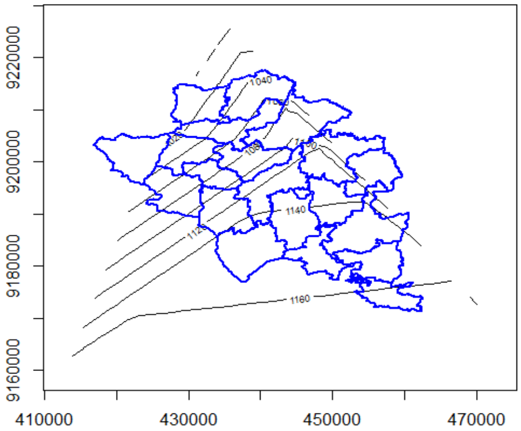


Fig. 2. Isohyet Map of Dry Seaon

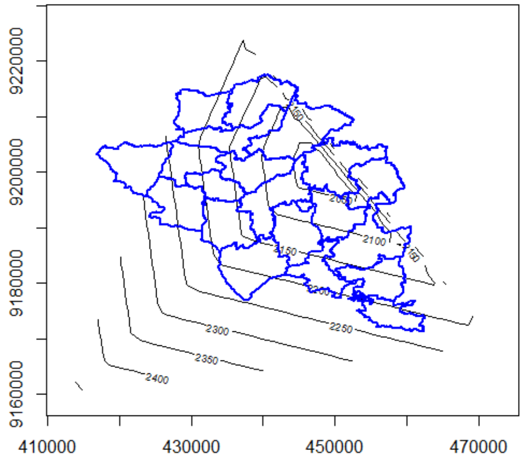


Fig. 3. Isohyet Map of Wet Season

For illustrating isohyet map in Semarang, it needs to figure out that the isohyet map [2] will be illustrated for dry season (April – September) and wet season (October – March) [4]. VAR model that was previously noted to forecast the rainfall in dry and wet season for each station was applied and the result is portrayed in Table 4 (the applied coordinate is UTM WGS84) [2]. In this case, the rainfall forecast for each of dry and wet season is the forecast cumulative number throughout the related season.

Isohyet map shown in Figure 2 and Figure 3 is the isohyet map for dry and wet season. The thick striped map is the map of Semarang with the boundaries of the existing sub-districts in the regency. The contour lines on the isohyet map portray the rainfall forecast associated with certain area (sub-district). By observing isohyet map as shown in Figure 2 and Figure 3, the decision makers can estimate/predict the rainfall that will drop in certain area (sub-district) in dry and wet season as well, hence the plant that will be planted in the related area can be determinedafterwards regarding the characteristic (water requirement) of the plant.

V. Conclusion

VAR (6) model can be applied well to forecast the rainfall in Semarang in dry and wet season. VAR (6) is used for the reason that through ADF calculation, each time series in the existing measurement stations is all stationary. Meanwhile, VAR (6) is taken as regarding the AIC calculation, this model has the lowest/smallest value than the other VAR models. VAR (6) can be applied well in Semarang as having the relatively small values of MAE and MAPE (valued below 10%) and smaller than ARIMA model used by BMKG (valued about 22,13%). Based on the mathematical model formulated with VAR (6) model based, isohyet map for each dry and wet season was made. It will be beneficial for the decision making stages as showing the correlation between certain areas (in this case – certain sub-districts in Semarang) and the rainfall (forecast) in the related areas.

References

[1] Abustan, Mahyuddin. (2009). *Analysis of vector autoregressive (VAR) to the correlation between public expenditure and economic growth in South Sulawesi*. Journal of Development Economics, vol. 10, no. 1, June 2009, pp. 1-14.

[2] Bivan, Roger S., Edzer J. Pebema, Virgillio Gomez, Rubio. (2008). *Applied spatial data analysis with R.* New York: SpringerScience+Business Media, LLC.

[3] Cowpertwait, Paul S.P., Andrew V. Metcalfe. (2009). *Introductory time series with R.* New York: Springer Science+Business Media, Inc.

[4] Damayanti, Noer Rochma, Muhammad Taufik, Eko Prasetyo, Parwati. (2008). *Isohyet mapping of Gerbang Kertasusila area based on NOAA-AVHR data.* Surabaya: Geomatics Engineering Program of FTSP ITS.

[5] Gujarati, Damodar N. (2006). *Essential of Econometrics.* New York: McGraw-Hill Co.

[6] Huda, Ary Miftakhul, Achmad Choiruddin, Osalliana Budiarto, Sutikno. (2012). *Rainfall data forecast using seasonal autoregression moving average (SARIMA) with outlier detection for agricultural production optimization effort in Mojokerto*. National Seminar of Food and Energy Sovereignty. Madura: Faculty of Agriculture, Trunojoyo University.

[7] Im, Kyung So, M. Hashim Pesharan, Yongcheol Shin. (2003). *Testing for unit roots in heterogeneous panels*.Journal of Econometrics, 115, pp. 53-74.

[8] Joshua. (2007). *Analysis of vector autoregression (VAR) to interrelationship of GDP growth and employment opportunity growth (A case study of Indonesia in 1977-2006),* University of Indonesia, Faculty of Mathematics and Natural Sciences, Department of Mathematics.

[9] Lutkepohl, Helmut. (2005). *New introduction to multiple time series analysis.* Berlin: Springer Science+Business Media, Inc.

[10] Mauriccio, Jose Alberto. (1999). *An algorithm for the exact likelihood of a stationary vector autoregression moving average*.Journal of Time Series Analysis, vol. 23, no. 4, ISSN 0143-9782/02/04, pp. 473-486.

[11] Okky, Dimas, Setiawan. (2012). *Composite stock price index modeling (IHSG), exchange rate and world oil price modeling using vector autoregression approach*. Journal of Science and Arts of ITS, vol. 1, no. 1.

[12] Raghunath, H.M. (2006). *Hydrology: Principles, analysis, design.* New Delhi: New Age International (P) Limited Publishers.

[13] Salvatore, Dominick, Derrick Reagle. (2002). *Theory and problems of statistics and econometrics.* New York: McGraw-Hill.

[14] Saputro, Dewi Retno Sari, Aji Hamim Wigena, Anik Djuraedah. (2011). *Autoregressive model for rainfall forecast in Indramayu*.Statistics and

Computing Forum, October 2011, vol. 16, no 2, pp.7-11. ISSN: 0853-8115.

[15] Schumway, Robert H., David S. Stoffer. (2011). *Time series analysis and its application.* New York: Springer Science+Business Media, Inc.

[16] Srinivasan, Shuba, Marc Vanhuele, Koen Pauwels. (2010). *Mind-set metrics in market response models: An integrative approach*.Journal of Marketing Research, August 2010, vol. 47, pp. 672-684.

[17] Widarjono, Agus. (2013). *Econometrics: Introduction and application with EViews guide*. Yogyakarta: UPP STIM YKPN.

[18] *Definition of contour map and isohyet*. Retrieved August 19, 2013 from http://www.hko.gov.hk/wxinfo/rainfall/isohyete.shtml

[19] *Geography, topography and geology of Semarang*. Retrieved July 10, 2013 from http://www.semarangkab.go.id/utama/selayang-pandang/kondisi-umum/geografi-topografi.html

[20] *Dry and wet season in Indonesia*. Retrieved July 11, 2013 from http://www.bmkg.go.id

[21] *Research and Development sites of agricultural commodity – Department of Agriculture*. Retrieved July 20, 2013 from http://bbsdlp.litbang.deptan.go.id/tamp_komoditas.php

[22] *Agriculture and plantation areas in Indonesia*. Retrieved July 25, 2013 from http://indonesia.go.id/en/potential/natural-resources

AUTHOR PROFIEL

Adi Nugroho earned a bachelor's degree from Geological Engineering of Bandung Institute of Technology (ITB), Indonesia and a master's degree majoring in Information System Management in Gunadarma University, Jakarta, Indonesia. He is currently finishing his doctoral program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia. He has interests in programming, databases, and software engineering. He is currently a lecturer in Faculty of Information Technology of Satya Wacana Christian University, Salatiga, Indonesia.

Subanar earned a bachelor's degree from Gadjah Mada University, Yogyakarta, Indonesia and a doctoral degree from Wisconsin University, United States. His field of research is statistics. He currently serves as a professor in Doctoral Program majoring in Computer Science inFacultyof Mathematics and Natural Sciences of Gadjah Mada University,Yogyakarta, Indonesia.

Sri Hartati earned a bachelor's degree from Department of Electronics and Instrumentation of Gadjah Mada University, Indonesia and a master's degree majoring in Computer Science (Spatial Processing) in New Brunswick University, Canada. Her doctoral degree was also earned from New Brunswick University, Canada majoring in Computer Science (Artificial Intelligence). She currently serves as a professor in Doctoral Program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia.

Khabib Mustofa earned a bachelor's and a master's degree fromDepartment of Computer Science of Gadjah Mada University, Indonesia and adoctoral degree from Vienna Universityof Technology, Austria majoring in Computer Science. His fields of research are Web Technology, Semantic Web, Software Engineering, and InformationManagement. He currently serves as a lecturer in Doctoral Program majoring in Computer Science in Faculty of Mathematics and Natural Sciences of Gadjah Mada University, Yogyakarta, Indonesia.

A Three-Dimensional Motion Anlaysis of Horse Rider in Wireless Sensor Network Environments

Jae-Neung Lee
Dept. of Control and Instrumentation Engineering,
Chosun University, 375 Seosuk-dong
Gwangju, Korea

Keun-Chang Kwak*
Dept. of Control and Instrumentation Engineering,
Chosun University, 375 Seosuk-dong
Gwangju, Korea

***Abstract*—This paper constructed a database of the national representative level of a professional horse-rider by wearing a motion-capture suit attached with 16 inertial sensors under an inertial sensor-based wireless network environment, then made a visual comparative analysis through a few methods (graphical and statistical) on the values of all motion features (elbow angle, knee angle, knee-elbow distance, backbone angle and hip position) classified depending on horse types (using two horses named Warm-blood and Thoroughbred) and footpace types (at a trot and a canter) and obtained by various methods of calculating Euclidean distance, the second cosine, maximum and minimum values, and made a comparative analysis depending on motion features of a horse-rider by using MVN studio software. In the study, the experimental results confirmed the validity of the proposed method of obtaining the motion feature database of a horse-rider in the wireless sensor network environment and making an analytical system.**

Keywords*—3D motion capture and analysis; inertial sensor; wireless network

I. Introduction

A lot of people are doing exercises to keep a good body shape. Especially, everybody knows that obesity is good neither for appearance nor for health. Horse-riding is a good sport of keeping good health and body line. It is possible to analyze and properly coach the postures of a horse-rider by the analysis of the so-called horse-riding motions under the wireless sensor network environment. Particularly, horse-riding is an exercise with a special trait that a horse-rider and a horse alive should be joined together. It can be helpful for a horse-rider to build up physical health and spiritual growth. In addition, the horse-riding is a physical exercise of the whole body helpful to improve body’s balance and flexibility for general physical developments, and it is also a spiritual exercise helpful to bring a spirit of boldness and sound thinking and to cultivate humanity through the learning process of being kind to animals. In summary of the horse-riding related mental, physical and psychological effects, it helps mentally to improve self-confidence through the talks with animals in love, learn social order in the process of horse-riding activities and cultivate patience. In addition, as it is an exercise of the whole body, it helps physically to improve blood circulation by using muscles and joints and general body adaptability and flexibility in relation to functional recovery, a sense of balance and a change in speed by stimulating all the nerves of several body parts. Furthermore, it has positive psychological impacts of having a satisfactory feeling about properly dealing with a big animal like a horse, a respectful mind for the dignity of life and an emotional growth in inter-personal relationships through enhancement of active attitudes in general. Finally, it is directly effective to correct body postures, enhance bowel function, enlarge lung capacity, prevent arthritis, anemia and constipation, build up courage, increase back flexibility, promote body rhythms and strength pelvis [1-3].

However, the horse-riding is not effective if correct body postures are not learned and maintained properly in the course of the exercise. Therefore, in order to make the most effective achievement within the shortest period of time, it is necessary to get a coaching session to check what is wrong with the motions of a horse-rider. Recently a lot of studies have been made by using inertial sensors at the wireless sensor network environment to carry out the most effective coaching session of the horse-riding sport. Luinge [4] proposed a method of accurately measuring the size of a person by using inertial sensors and angular velocity sensors.

Zhou [5] suggested a new human motion tracking system using two wearable inertial sensors that are placed near the wrist and elbow joints of the upper limb. Lee[6] suggested Sensor fusion and calibration for motion captures using accelerometers. Zhu [7] presented a real time motion-tracking system using tri-axis micro electromechanical accelerometers. Cheng [8] suggested the results of a set of network traffic experiments that were designed to investigate the suitability of conventional wireless motion sensing system design which generally assumes in-network processing - as an efficient and scalable design for use in sports training. Venkatraman [9] analyzed a behavior of the animal is further extracted from the recorded acceleration data using neural network based pattern recognition algorithms. Ghasemzadeh [10] suggested a golf swing training system which incorporates wearable motion sensors to obtain inertial information and provide feedback on the quality of movements. Kevin [11] analyzed a theory, design, and evaluation of a miniature, wireless IMU (Initial Measurement Unit) that precisely measures the dynamics of a golf club used in putting. Mariani [12] described 3D gait assessment in young and elderly subjects using foot-worn inertial sensors. Yujin[13] suggested Upper Body Motion Tracking With Inertial Sensors. Lijun [14] analyzed A Practical Calibration Method on MEMS Gyroscope. Wei [15] suggested calibration of low-precision MEMS inertial sensor. Cao [16] performed 3D dynamics analysis of a golf full swing by fusion inertial sensors and vision data. Jung [17] analyzed smart shoes. Chan [18] proposed a virtual reality dance training system using motion capture technology. Frosio [19] analyzed automatic calibration of MEMS accelerometers.

However, a variety of inertial sensor-based studies have been made so far, but no study has been made about an analysis on the motions of a horse-rider in a wireless network environment. Thus, in this paper, a database is constructed by collecting the motions of a professional horse-rider wearing a motion-capture suit under the inertial sensor-based wireless network environment. At this time, two representative types of horses, named Warm-blood and Thoroughbred, were selected for this study. Actual horse-riding sessions were made to analyze the postures of a professional horse-rider by measuring and calculating all the motion feature values of elbow angle, knee angle, knee-elbow distance, backbone angle and hip y-axis position at the two representative horse-riding footpace types, rising trot and canter.

The results of the experiments made at MVN Studio with MATLAB confirmed the validity of the method of analyzing the motions of a horse-rider in a wireless sensor network environment suggested in this study, in which all the motion data of a horse-rider were collected in the wireless sensor network environment [20] to make a comparative analysis on all the horse-riding postures carefully.

II. Method of Constructing an Inertial Sensor-Based Wireless Network Environment and a Database

This chapter describes a method of constructing an inertial sensor-based wireless network environment and a motion database. All the motion data of a horse-rider are received to a computer through the MVN motion capture system constructed with inertial sensors made by Xsens Co. Then, all the data are compared by using respective calculation methods. Figure 2 below illustrated the steps of collecting and constructing a database of motions in a wireless sensor network environment. Differently from an optical sensor-based motion capture system, the MVN motion capture system can capture the entire body motions wirelessly without using a camera. In addition, the MVN motion capture system is portable for convenient indoor-outdoor uses.

Figure 1 illustrates a process of wireless sensor network made by Xsens Co. The MTX sensor used for constructing the database is a small and light 3DOF Human orientation tracker which provides drift-free kinematic data, precise three-axis acceleration, three-axis gyroscope and three-axis geomagnetic values. Figure 3 shows the sensor used for constructing a database.

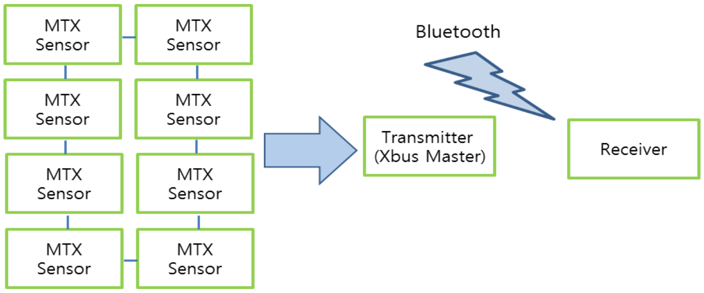


Fig. 1. Process of wireless sensor network

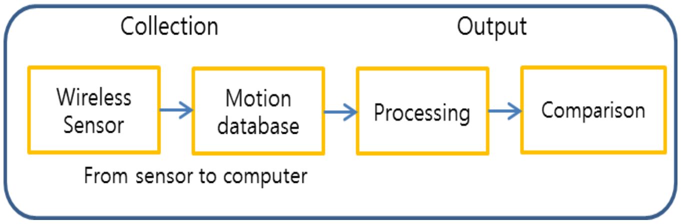


Fig. 2. Data collection and output processing

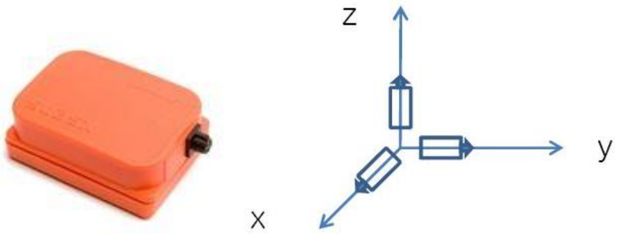


Fig. 3. MTX sensor: 3DOF human orientation tracker

In order to collect data, this study used a subject, a national representative level of a professional horse-rider whose height is 164cm with her foot size of 235mm. She worn the MVN, the inertial sensor-based 3D motion capture suit made by Xsens, to collect data, subsequently riding on the horses, thoroughbred whose height is160cm and Warm-blood whose height is 150~173cm. The period of time taken for measurement of one file was about 1~2 minutes and 15 data were collected depending on footpace types. There are 4 horse footpace types such as walk, trot, canter and gallop. A horse usually goes as far as 130m for a minute, approximately 8km for an hour at a walk. It usually moves as far as 220m for a minute, approximately 13km for an hour at a rising trot, one specific type of trots. It generally moves as far as 350m for a minute, approximately 21km for an hour at a canter. It moves as far as 1000m for a minute, approximately 60km or even 72km for an hour at a gallop. The horses used in the experiments of this study were made to move at the two footpace types, a rising trot and a canter. The measurement frame rate was 100 frames per second. Figure 4 (a) show the two horse types respectively, Warm-blood and Thoroughbred, on which the horse-rider was sitting, in the process of constructing the actual DB. Figure 4 (b) illustrates all the 16 body parts attached with inertial sensors, differently marked for her visible front and lateral sides and her invisible back side.

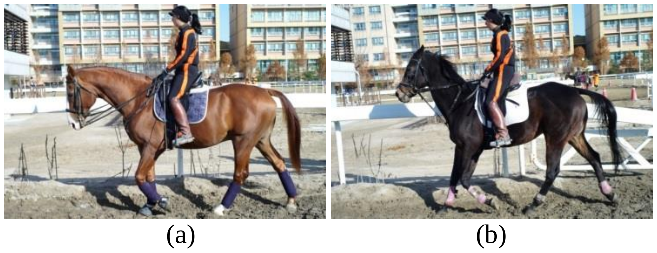

(a) (b)

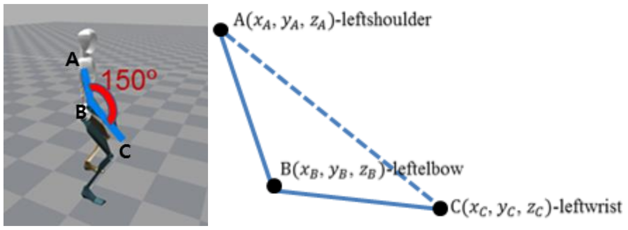
Fig. 4. (a) Warm-blood of Actual horse-rider's postures (b) Thoroughbred of Actual horse-rider's postures

III. Method of Extracting Features from Motions of Horse-Rider

If body postures are not properly learned and maintained by a horse-rider, the horse-riding exercise may bring not a positive, but a negative effect. Therefore, it is necessary to make an analysis on the horse-riding postures to clearly check what is wrong with the motions of the horse-rider. The following method is presented to make a comparative analysis on 5 motion features (elbow angle, knee angle, elbow-knee distance, backbone angle, hip position).

A. Elbow Angle

The values of motion features collected by the sensors attached at body parts, A(wrist), B(elbow) and C(shoulder), are extracted from the DB to define three coordinates, A, B and C through Eq. (1). Figure 5 (a) shows MVN studio motion capture software. Figure 5 (b) illustrates a method of obtaining an elbow angle.


A
150°
B
C
$A(x_A, y_A, z_A)$-leftshoulder
$B(x_B, y_B, z_B)$-leftelbow
$C(x_C, y_C, z_C)$-leftwrist

(a) (b)

Fig. 5. (a) MVN studio software (b) Method of obtaining an elbow angle

$$A = x_A, y_A, z_A \ , \ B = x_B, y_B, z_B \ , \ C = x_C, y_C, z_C \qquad (1)$$

Besides, all the respective distances among feature positions, A(wrist), B(elbow), C(shoulder), are obtained in Euclidean geometry through Eq.(2)

$$\overline{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2)} = c$$

$$\overline{BC} = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2 + (z_B - z_C)^2)} = a$$

$$\overline{CA} = \sqrt{(x_C - x_A)^2 + (y_C - y_A)^2 + (z_C - z_A)^2)} = b \qquad (2)$$

The respective distances among the feature positions are obtained and applied to the following Eq.(3).

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc\cos A \\ b^2 &= c^2 + a^2 - 2ca\cos B \\ c^2 &= a^2 + b^2 - 2ab\cos C \end{aligned} \qquad (3)$$

A change is made into Eq.(4), so as to calculate an elbow angle.

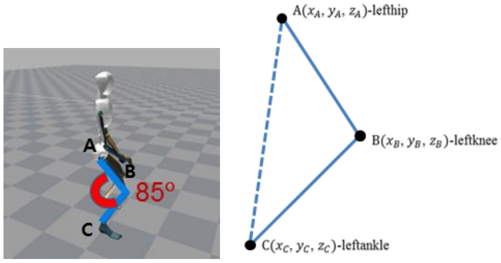
$$\text{elbow angle} = cos^{-1}(\frac{c^2 + a^2 - b^2}{2ca}) \qquad (4)$$

B. Knee Angle

The values of motion features collected by the sensors attached at body parts, A(left hip), B(left knee) and C(left ankle), are extracted from the DB to define three coordinates, A, B and C in the same way shown in case of the elbow angle. The distances among feature positions, A(left hip), B(left knee), C(left ankle), are obtained in Euclidean geometry. The respective distances among the feature positions are obtained and then applied to Eq. (3). A change is made into Eq. (5), so as to calculate the knee angle.

$$\text{knee angle} = cos^{-1}(\frac{c^2 + a^2 - b^2}{2ca}) \qquad (5)$$

Figure 6 (a) shows MVN studio motion capture software. Figure 6 (b) illustrates a method of obtaining a knee angle.


A
B
85°
C
$A(x_A, y_A, z_A)$-lefthip
$B(x_B, y_B, z_B)$-leftknee
$C(x_C, y_C, z_C)$-leftankle

(a) (b)

Fig. 6. (a) MVN studio software (b) Method of obtaining a knee angle

C. Elbow-Knee Distance

The values of motion features collected by the sensors attached at body parts, A(left elbow) and B(right elbow), are extracted from the DB to define two coordinates, A and B as shown in Figure 6 (b). An elbow distance between the features A and B is obtained through Eq. (6).

$$\text{Elbow distance} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2)} \qquad (6)$$

In the same way described above, the values of body features collected by the sensors attached at body parts, C(left knee) and D(right knee), are extracted from the DB to define two coordinates, C and D as shown in Figure 7. A knee distance between the features C and D is obtained through Eq. (7)

$$\text{knee distance} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2)} \qquad (7)$$

Figure 7 illustrate methods of obtaining an elbow distance and a knee distance, respectively.

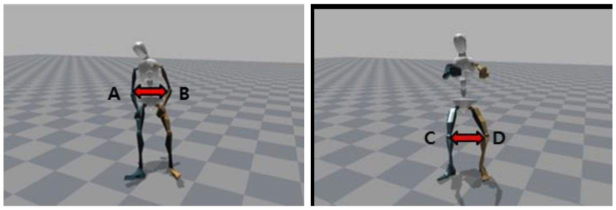


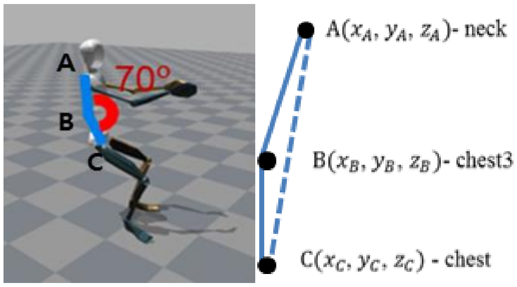
Fig. 7. Method of obtaining an elbow and knee distance

D. Backbone Angle

The values of motion features collected by the sensors attached at body parts, A(neck), B(chest) and C(chest), are extracted from the DB to define three coordinates, A, B and C in the same way shown above in case of the elbow angle. The distances among feature positions, A(neck), B(chest3) and C(chest), are obtained in Euclidean geometry. The respective distances among the feature positions are obtained and then applied to Eq. (3). A change is made into Eq. (8), so as to calculate a backbone angle.

$$\text{backbone angle} = cos^{-1}(\frac{c^2 + a^2 - b^2}{2ca}) \tag{8}$$

Figure 8 (a) shows MVN studio motion capture software. Figure 8 (b) illustrates a method of obtaining a backbone angle.



(a) (b)

Fig. 8. (a) MVN studio software (b) Method of obtaining a backbone angle

E. Hip Position

The value of a body feature collected by the sensor attached at a body part, A(hip), is extracted from the DB to define a coordinate, A(x,y,z). As the hip position is centered at horse-riding, the movement of A(y) axis is in use.

$$\text{Hip Position} = \text{A}(Y_A) \tag{16}$$

IV. Experimental Results

In the experiments of this study, data were collected by making two horse types (Warm-blood and Thoroughbred) at two footpace types (rising trot and canter) 15 times, so that 4 cycles of data were extracted for comparative analysis. It was confirmed that there was differences in horse-rider's postures through different values of respective angles and distances. The right elbow and knee are marked in red dotted lines.

A. Warm-blood at a Trot

The following figure shows that repeating 200 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (2.5 seconds). As shown in the Figure below, the elbow and knee angles remain at the range of about 130~150 degrees and about 130-170 degrees, respectively. The elbow and knee distances stay at the range of 18~27cm and 15~18cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 30~37cm. Figure 9 shows the average feature values of 15 data.

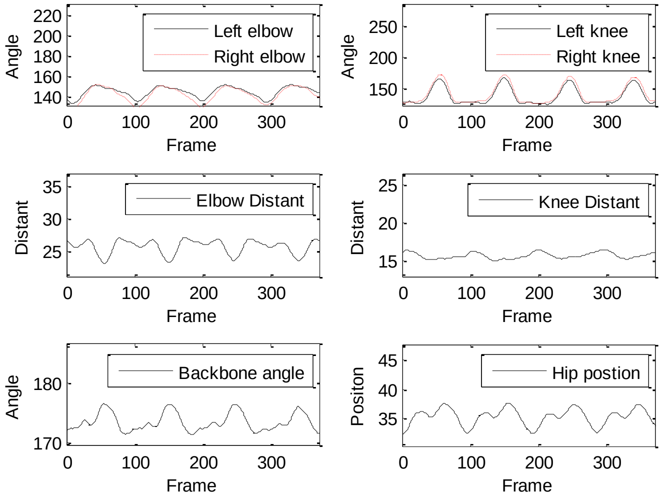


Fig. 9. Feature values of Warm-blood at a trot

B. Warm-blood at a Canter

Figure 10 shows that repeating 300 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (3 seconds). As shown in the Figure below, the elbow and knee angles remain at the range of about 130~160 degrees and 130~140 degrees, respectively. The elbow and knee distances stay at the range of 25~29cm and 15~18cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 30~37cm. Figure 10 shows the average feature values of 15 data.

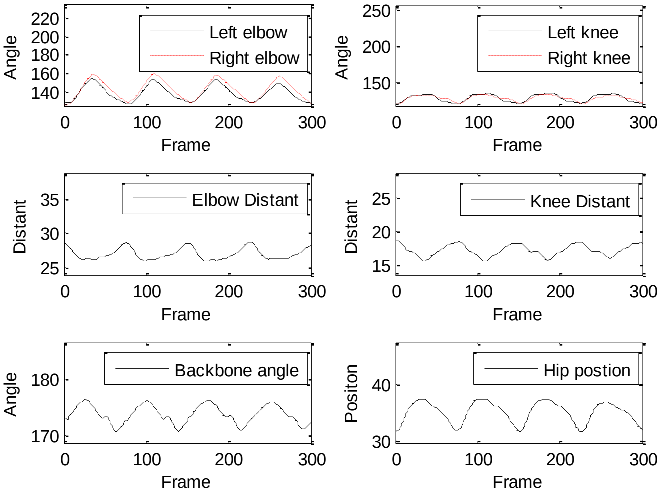


Fig. 10. Feature values of Warm-blood at a canter

C. Thoroughbred at a Trot

Figure 11 shows that repeating 200 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (2.5 seconds). As shown in the Figure below, the elbow and knee angles remain at the range of about 135~160 degrees and about 120~160 degrees, respectively.

The elbow and knee distances stay at the range of 22~25cm and 14~16cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 34~40cm. Figure 11 shows the average feature values of 15 data.

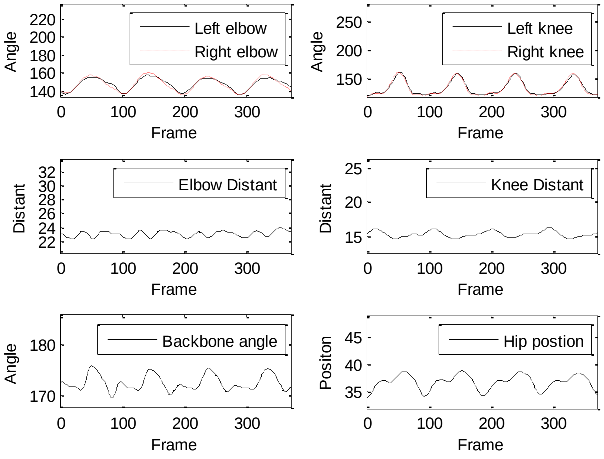


Fig. 11. Feature values of Thoroughbred at a trot

D. Thoroughbred at a Canter

Figure 12 shows that repeating 300 frames were extracted out of approximately 10000 frames to demonstrate 4 cycles (3 seconds). As shown in the Figure below, the elbow and knee angles remain at the range of about 135~160 degrees and about 120~140 degrees, respectively. The elbow and knee distances stay at the range of 22~25cm and 13~19cm, respectively. The backbone angle remains at the range of 170~177 degrees. The Hip positions moves within the range of 31~38cm. Figure 12 shows the average feature values of 15 data.

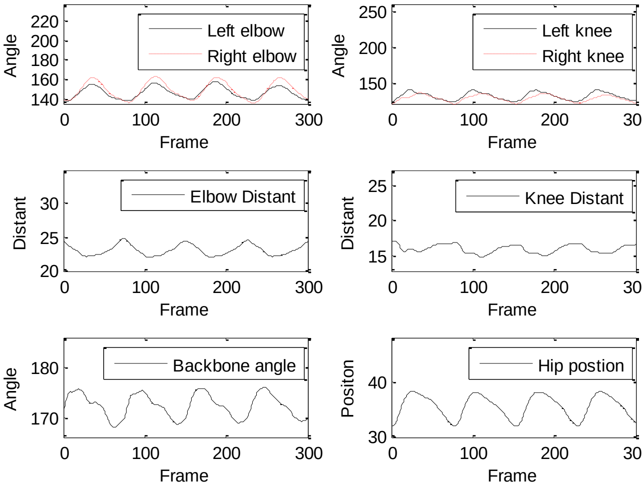


Fig. 12. Feature values of Thoroughbred at a canter

Figure 13 below illustrates the numerical comparison of feature values of two horses at a rising trot. A visible difference is revealed in the elbow angles. A similar difference is noticed in knee angles as shown in the elbow angles. The reason why no significant difference was made in the backbone angles and in the hip positions is because the horse-rider should keep her backbone at its perpendicularity and her hip at the same position whatever type of a horse she is riding on.

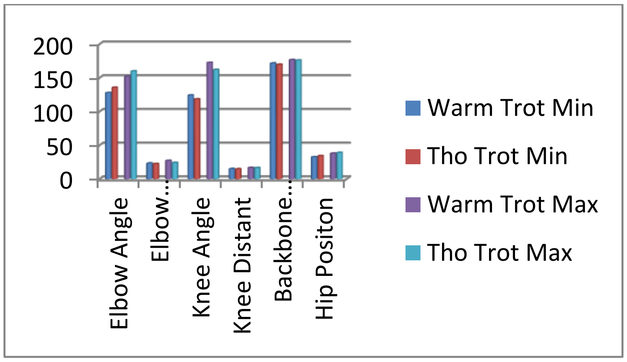


Fig. 13. Comparison of maximum and minimum feature values at a trot

Figure 14 below illustrates the numerical comparison of maximum and minimum feature values of two horses (Warm-blood and Thoroughbred) at a canter. A visible difference is revealed in the elbow angles. A similar difference is noticed in knee angles as shown in the elbow angles. As described above, no significant difference was made in the backbone angles and in the hip positions because the horse-rider should keep her backbone at its perpendicularity and her hip at the same position regardless of the type of horses.

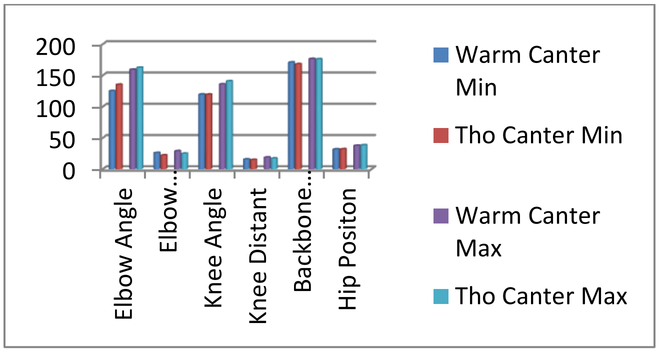


Fig. 14. Comparison of maximum and minimum feature values at a canter

V. Conclusions

This paper suggested a method of using a motion database of a professional horse-rider wearing a suit constructed with wireless networks consisting of 16 inertial sensors and then extracting the respective motion features (elbow angle, knee angle, backbone angle, hip position, knee-elbow distance) through various calculation methods such as Euclidean distance, the second cosine, maximum and minimum values, depending on horse types (Warm-blood and Thoroughbred) and footpace types (trot and canter). MVN studio software was used to make a comparative analysis on the horse-rider's motion features depending on the footpace types. As a result, a significant difference was noticed in the motion feature values obtained depending on different horse and footpace types. Therefore, in order to effectively make real-time coaching sessions for different horse-riding footpace types, it is necessary to construct a motion feature database in relation to footpace types and accordingly make a suitable analysis and coach on horse-riding motions.

References

[1] The MTx system, xSens, http://www.xsens.com/

[2] Horse riding, http://terms.naver.com/entry.nhn?docId=384601&cid=689&categoryId=1458

[3] How to Horse riding, http://ko.wikipedia.org/wiki%EX%8A%B9%EB%A7%88

[4] H. J. Luinge, and P. H. Veltink, “Measuring Orientation of Human Body Segments Using Miniature Gyroscopes and Accelerometers,” Medical & Biological Engineering & Computing, vol. 43, no. 2, pp. 273-282, 2005.

[5] H. Zhou, T. Stone, H. Hu, and N. Harris, “Use of Multiple Wearable Inertial Sensors in Upper Limb Motion Tracking,” Medical Engineering & Physics, vol. 30, no. 1, pp. 123-133, 2008.

[6] J. Lee, and I. Ha, “Sensor fusion and calibration for motion captures using accelerometers,” Proc. IEEE International Conference on Robotics and Automation, 1999, pp. 1954-1959.

[7] R. Zhu, and Z. Zhou, “A Real-Time Articulated Human Motion Tracking Using Tri-Axis Inertial/Magnetic Sensors Package,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 12, no. 2, pp. 295-302, 2004.

[8] L. Cheng, S. Hailes, “An Experimental Study on a Motion Sensing System for Sports Training,” short paper in the Proceedings of the 5th European Conference on Wireless Sensor Networks (EWSN), Bologna, Italy, Feb 2008.

[9] S. Venkatraman, J. Long, K. Pister, J. Carmena, “Wireless Inertial Sensors for Monitoring Animal Behaviour,” in Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBS), Lyon, France, Aug 2007, pp. 378-381.

[10] H. Ghasemzadeh, V. Loseu, E. Guenterberg, and R. Jafari, “Sport training using body sensor networks: a statistical approach to measure wrist rotation for golf swing,” in BodyNets ’09: Proceedings of the Fourth International Conference on Body Area Networks, 2009, pp. 1–8.

[11] K. Kevin, S. W. Yoon, N.C. Perkins, “Wireless mems inertial sensor system for golf swing dynamics,” Sensors and Actuators A: Physical, vol. 141, no. 2, pp. 619 – 630, 2008.

[12] Mariani B, Hoskovec C, Rochat S, Bula C, Penders J, Aminian K. “3D gait assessment in young and elderly subjects using foot-worn inertial sensors,” J Biomech, 2010;43(15), pp.2999–3006.

[13] J. Yujin, K. Donghoon, K. Jinwook, “Upper Body Motion Tracking With Inertial Sensors,” proc. IEEE International Conference on Robotics and Biomimetics, 2010, pp. 1746-1751.

[14] Song Lijun, Qin Yongyuan, “A Practical Calibration Method on MEMS Gyroscope”. Piezoelectrics & Acoustooptics, Vol.32, No.3, 2010, pp.372-374

[15] R. Wei, Z. Tao, Z haiyun, W. Leigang, “A Research on Calibration of Low-Precision MEMS Inertial Sensors,” IEEE Transactions on Control and Decision Conference, 2013, pp. 3243-3247.

[16] N. Cao, S. Young, K. Dang, “3D Dynamics Analysis of a Golf Full Swing by Fusing Inertial Sensor and Vision data,” International Conference on Control, Automation and Systems, 2013, pp. 1300-1303.

[17] P. G. Jung, G. Lim, K. Kong, “A Mobile Motion Capture System Based on Inertial Sensors and Smart Shoes,” IEEE Transactions Robotics and Automation, 2013, pp. 692-697.

[18] J. C. P. Chan, H. Leung, J. K. T. Tang, and T. Komura, “A virtual reality dance training system using motion capture technology,” IEEE Transactions on Learning Technologies, vol. 4, no. 2, pp. 187.195, 2011.

[19] [19] I. Frosio, F. Pedersini, and N. A. Borghese, “Autocalibration of MEMS accelerometers,” *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 6, pp. 2034–2041, 2009.

[20] Wireless sensor network, http://en.wikipedia.org/wiki/ Wireless_sensor_network

Authors Profile

Jae-Neung Lee received the B.Sc. from Chosun University, Gwangju, Korea, in 2013. He is currently pursuing a candidate for the M.Sc. His research interests include human–robot interaction, computational intelligence, and pattern recognition.

Keun-Chang Kwak received the B.Sc., M.Sc., and Ph.D. degrees from Chungbuk National University, Cheongju, Korea, in 1996, 1998, and 2002, respectively. During 2003–2005, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. From 2005 to 2007, he was a Senior Researcher with the Human–Robot Interaction Team, Intelligent Robot Division, Electronics and Telecommunications Research Institute, Daejeon, Korea. He is currently the Associative Professor with the Department of Control and Instrumentation, Engineering, Chosun University, Gwangju, Korea. His research interests include human–robot interaction, computational intelligence, biometrics, and pattern recognition. Dr. Kwak is a member of IEEE, IEICE, KFIS, KRS, ICROS, KIPS, and IEEK.

Data and Knowledge Extraction Based on Structure Analysis of Homogeneous Websites

Mohammed Abdullah Hassan Al-Hagery
Qassim University, Faculty of Computer, Department of IT
Buraydah, KSA

***Abstract*—The World Wide Web includes several types of website applications. Mainly these applications are related to business, organizations, companies, and others. There is a lack to get raw data sets to study the behavior of the internal structure of each type of these websites. Where websites structures include treasure of links, and sub-links, in addition to some embedded features associated with the internal structure of each website. The objective of this paper is to analysis a set of homogeneous websites to establish raw data sets. These data sets can be employed for several research purposes. It also can be used to extract some invisible aspects/features within the structure. Several steps are required to accomplish this objective; first, to propose an algorithm for structure analysis, second, to implement the proposed algorithm as a software tool for the purpose of extraction and establishment of raw data sets (real data set), third, to extrapolate a set of rules or relations from these data sets. This data set can be employed for researches purposes in the field of web structure mining, to estimate important factors related to websites development processes, and websites ranking. The results comprise creation of Oriented Data Sets (ODS) for research purposes and also for deducing a set of features represents a type of new discovered knowledge in this ODS.**

Keywords*—Hyperlinks Analysis Tools; Features Extraction; Oriented Data Sets generation; Knowledge Discovery in Oriented Data Sets

I. Introduction

The internet is becoming a main communication tool between various people, companies, and organizations in society. It has become the dominant force in the world in various areas where the information revolution represents the strength of economy and uplifts the level of people’s life. In addition, the internet has entered many practical areas in our life such as business management, purchase operation, follow up economy of knowledge stocks, trading prices, currencies index, banking services, E-governments, distance learning, hospitals, and other areas. The internet websites contain huge amount of data. It includes simple and complex data within web links. Although there are many companies and organizations established some software tools and methods used to analyze these web links or web contents. Most of them are focusing on particular attributes, relevant to a specific objective. All these tools cannot be employed to collect the required data for this research.

This research is leading to a new idea focusing mainly on the analysis of web structure components that depend on the dynamic hyperlinks. This task is used to extract different useful features of a website. Few numbers of these features will be applied as needed in this research, for example, in estimation of websites' sizes or in ranking processes relevant to websites structure. The extracted data and features in this paper will be directed for serving websites development process as an extension for this research.

II. Literature Review

There are many researches and studies that focus on web structure analysis and web structure mining, but with different objectives. All these works are based on using many tools and methods that serve the goals of such research and studies. Some of these works will be discussed here. Many research works have been undertaken and different solutions have been suggested to the problem of searching, indexing or querying the web, taking into account its structure as well as the meta-information included in the hyperlinks and the text surrounding them [10], [5], [12] and [16]. There are a number of algorithms proposed based on link analysis. Dean and Henzinger [7] proposed an algorithm to exploit only the hyperlink-structure (i.e. graph connectivity) of any Website and does not examine the information about the content or usage of pages or structure components. Brin and Page [4], addressed the question of how to build a practical large-scale system which can exploit the additional information present in hypertext. They aimed to speed up Google considerably through hardware distribution, software, and algorithmic improvements. Their target was to handle the several hundred queries per second.

Jon and Kleinberg [15], developed a set of algorithmic tools for extracting information from hyper link structures from web environments. They did some experiments that demonstrated their effectiveness in a variety of contexts on the WWW. The central issue they addressed within their framework was the distillation of broad search topics, through the discovery of “authoritative" information sources on such topics. They proposed and tested an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of “hub" pages" that joined them together in the link structure.

Xing and Ghorbani developed a Weighted PageRank (WPR) algorithm. It was an extension to the standard PageRank algorithm. The developed algorithm takes into account the importance of both the in-links and the out-links of the pages and distributes rank scores based on the popularity of the pages. The results showed that the WPR performs better than the conventional PageRank algorithm in terms of returning larger number of relevant pages to a given query [20].

Taherizadeh and Moghadam proposed an approach to integrate web content mining into web usage mining. The textual content of web pages is captured through extraction of frequent word sequences, which are combined with web server log files to discover useful information and association rules about users' behaviors [19].

Kao and Lin have proposed an algorithm called DRank to diminish the bias of PageRank-like link analysis algorithm that attains better performance than Page Quality. In their algorithm, they modeled web graph as a three-layer graph which includes Host Graph, Directory Graph, and Page Graph by using the hierarchical structure of URLs and the structure of link relation of Web pages [14]. In addition, Kumar and Singh introduced a study on hyperlink analysis. They analyzed the links in order to retrieve web information. They used Google search engine and different algorithms for link analysis, such as PageRank, Weighted PageRank, and Hyperlink-Induced Topic Search algorithms [17].

Jeyalatha and Vijayakumar [13] proposed and implemented a web link extraction tool to deal with web structure using Java and standard interface. They used the Breadth First Search strategy. This work is mainly focusing on performing a quick check on search links, analyzes the structure information from the web that includes document structure & hyperlinks, to Crawl HTML files, and counts the number of occurrences of the keywords in those files. The research helps web users, faculty, students and Web administrators in a university environment.

Mishra, et al. introduced their work based on PageRank created to rank the results of a search system based on a user's topic or query. More than one algorithm was proposed in their work [18]. Derouiche et al. in [8], presented a novel approach for extracting structured data from websites, and the goal was to harvest real-world items from the structured web. They proposed an alternative approach to automatic information extraction and integration from structured Web pages.

Based on the topology of the hyperlinks, Web Structure mining categorizes the web pages and generates the information like similarity and relationship between different web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level. It is important to understand the web data structure for information retrieval [18]. The web structure analysis can be performed by several ways, based on the type of required information or knowledge, then creating or using appropriate software tools to accomplish the analysis task. Some tools were employed to identify the relationship among web pages based on their contents or direct link connection, and other tools were created for different objectives. Birla et al., reported that the web is a treasure of information and data, where large amounts of data are available in different formats and structures. Finding the useful data from the web is a complex task [3].

Some of features extraction techniques are based on extracting content based features. However, many such solutions have been handcrafted and thus not guaranteed to work optimally under all data environments. Anand in his research explored an evolutionary algorithm based feature extraction techniques. This work explores Evolutionary algorithms based feature extraction techniques where the extracted features are used to describe user or item profiles [1]. On the other hand, Benslimane et al. proposed in their research idea a novel approach for reverse engineering data-intensive web application into ontology-based semantic web. They analyzed the HTML pages structure to identify its components, interrelationships, and extract a form model schema [2]. As discussed above and although, there is a number of tools and crawlers which can be used to analyze the websites, but it seems clear that the collected data sets by these tools were focusing to solve specific problems, not for everything we need, so one may not be able to take advantage of them, for example in the field of prediction and websites development and ranking as it is in the data sets of this search.

III. Websites Structure Anlysis

Web requirements include three classes: functional requirements, non-functional requirements, and other requirements. The traditional information retrieval system focuses on information provided by the text of web documents. Web mining technique provides additional information through hyperlinks where different documents are connected. The web may be viewed as a directed labeled graph where nodes are the documents or pages and the edges are the hyperlinks between them. The directed graph structure in the web is called as web graph [11]. A web can be imagined as a large graph containing several hundred million or billions of nodes or vertices and a few billion arcs or edges [17]. Link mining is divided into four parts; external structure mining, internal structure mining, URL mining, and web usage mining [6]. The structure analysis can be applied in several areas, such as query ranking, webpages importance, pages classification, and clustering. The objective of these types of analysis is to find the most related pages, redundancies, and measuring the similarity degree among pages.

IV. Problem Statement

There are many important features within the web structure that are invisible. This research is mainly focusing on the structure analysis of homogeneous websites based on link analysis, as a step forward for web structure mining. The challenge for this type is to deal with the structure of hyperlinks within the web itself. Link analysis is an old area of research. However, with the growing interest in web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [9], [6]. There is a difficulty to find a tool to produce the appropriate data relevant to the website hyperlinks contents and structure. Although there are many algorithms used to analyze the web links, these algorithms cannot be used to collect the required data for a specific objective, as in the objective of this research.

V. Objectives

The objective of this research is to propose and implement an algorithm in order to analyze hyperlinks of many homogeneous websites, to collect ODS to be used for research purposes, and also to extract some hidden features. The extracted features/rules can be used to help websites' developers to employ the results of this research in measurement and estimation models relevant to websites domain.

VI. Research Methodology

The analysis process concentrates on follow-up of dynamic hyperlinks of each website in order to discover interconnected links to find some important features. The research steps include; algorithm design, implementation, links analysis, generation of ODS, and features extraction.

A. *The proposed algorithm*

The proposed algorithm is designed to process websites' hyperlinks and to extract the required ODS & embedded features. Each website has its own structure and attributes values. The analysis results of hyperlinks are different from one website to another. The ODS have established based on the proposed algorithm. It has constructed for several homogeneous websites including many attributes. The Pseudo-codes of this algorithm are illustrated in the following two segments.

```
Start
Identify a list Li of Homogeneous websites links (HWL),
Li_length := N, Li={R1, R2,… RN} // N is size sample
Create ODS  with size N;
Define  a set of oriented attributes to be extracted from each Page
For each web  Page i= 1;  i > 0 and i ≤ N   do
Begin
    Read First Link(Ri);
    Initlize (Oriented_attrib_Record);
    S_Analysis (Ri , Oriented_attrib_Record); //  Proc- call
    Save (Oriented_attrib_Record, ODS[i]);
End;
Finish
```

The proposed algorithm can be used to extract the same attributes of any type of website not only the homogeneous websites.

```
S_analysis(Root_R:Link, Data_Rec: record): Record;
If (Current_Root_link=end of last branch   then exit
If (Current_Root_link<>Terminal leave OR  external Link)
Begin
   Scan all possible pages  connected  with Current_Root_link
   Calculate the attributes of the Current Page
   Update  the data record
   If there are links extensions for the current Page then
   Begin
        J:= the number of sub links  of Current_Root_link;
        Repeat
           Get (Link j);
           S_Analysis(Link j , Data_rec);  // Recursive Call
           J:= J -1;
      Until (j<=0)
   End  // if
End
Refresh  & Return(Data_rec);
Finish
```

B. *Web Analysis Processes*

The Web Analysis task includes many steps as shown in Figure 1. The steps are started by a user who enters the main hyper link/root of a website. The tool receives the root as an input and starts the analysis. It follows-up the sub links to the depth of the website in several directions and extracts the required data from all internal links. The extraction process stops when the following of current link is external link or final leaves.

The algorithm listed above illustrates the process of following-up the links. These algorithms developed as S/W tools that include several steps. These tools implemented by the PHP programming language based on its facilities enabling the tool to go to the depth of the website and to follow-up the discovery process of all hyperlinks. It is able to deal with web structure components and its contents. The main challenge of this research from the first step is to get the required data sets, for the purpose of showing this work into presence and leading it to success. The research data set is established and organized through several steps as real data sets. The proposed algorithm and the designed tools were essentially developed based on the needs of this research to achieve the desired objectives (web structure analysis, raw data establishment, and extraction of a special type of knowledge). The PHP language was applied through the local host XAMP. It is a suitable software environment for this work.

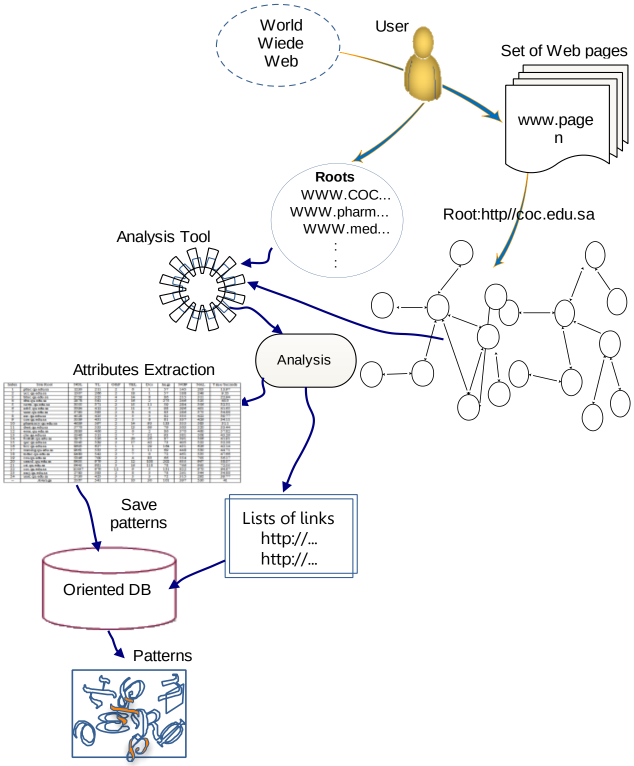


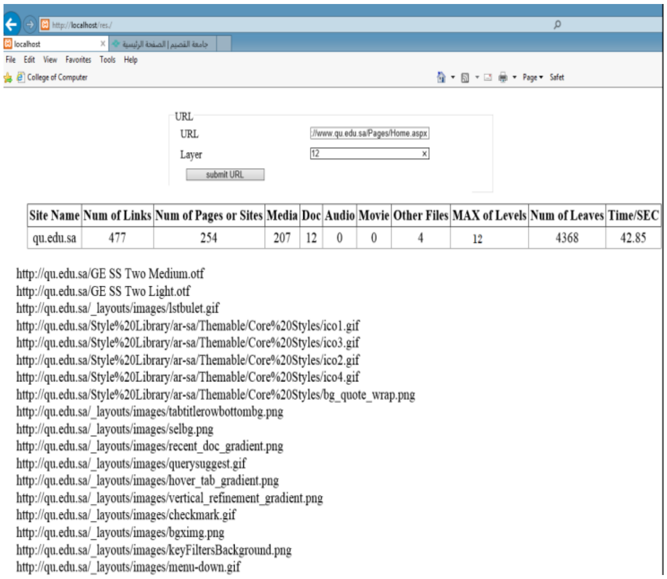
Fig. 1. Websites structure analysis

VII. Research Results

There are two types of results; first, set of raw data, denoted by ODS. Second, set of embedded features or rules. The following two sections A and B obtain more details about extracted results.

A. *Creation of raw Data sets*

Raw data sets consist of several attributes related to each website structure, such as, Total links, External Links, Number of Leaves, Active Links, No of Pages, Images, Docs, Other Files, Analysis Time/seconds, and etc. as presented in Table I and in Figure 2, these attributes were employed for new features extraction. The ODS extracted from 24 websites related to educational field from the websites of Qassim University. The contents are shown in Table I. This sample includes the following attributes; Site Root, Total number of Leaves (NOL), Total number of links (TL), Total number of External Links (TEL), Total number of Active Links (NAL), Total number of Pages (NOP), Images (Imgs), Docs (Dcs), Other Files (OthF), and analysis Time estimated in seconds. Figure 2 shows a part of the analysis results of one website at Qassim University.



URL

URL ://www.qu.edu.sa/Pages/Home.aspx

Layer 12

submit URL

Site Name	Num of Links	Num of Pages or Sites	Media	Doc	Audio	Movie	Other Files	MAX of Levels	Num of Leaves	Time/SEC
qu.edu.sa	477	254	207	12	0	0	4	12	4368	42.85

http://qu.edu.sa/GE SS Two Medium.otf
http://qu.edu.sa/GE SS Two Light.otf
http://qu.edu.sa/_layouts/images/lstbulet.gif
http://qu.edu.sa/Style%20Library/ar-sa/Themable/Core%20Styles/ico1.gif
http://qu.edu.sa/Style%20Library/ar-sa/Themable/Core%20Styles/ico3.gif
http://qu.edu.sa/Style%20Library/ar-sa/Themable/Core%20Styles/ico2.gif
http://qu.edu.sa/Style%20Library/ar-sa/Themable/Core%20Styles/ico4.gif
http://qu.edu.sa/Style%20Library/ar-sa/Themable/Core%20Styles/bg_quote_wrap.png
http://qu.edu.sa/_layouts/images/tabtitlerowbottombg.png
http://qu.edu.sa/_layouts/images/selbg.png
http://qu.edu.sa/_layouts/images/recent_doc_gradient.png
http://qu.edu.sa/_layouts/images/querysuggest.gif
http://qu.edu.sa/_layouts/images/hover_tab_gradient.png
http://qu.edu.sa/_layouts/images/vertical_refinement_gradient.png
http://qu.edu.sa/_layouts/images/checkmark.gif
http://qu.edu.sa/_layouts/images/bgximg.png
http://qu.edu.sa/_layouts/images/keyFiltersBackground.png
http://qu.edu.sa/_layouts/images/menu-down.gif

Fig. 2. Some extracted attributes of a single website

TABLE I. A Set of Extracted Attributes of 24 QU Colleges

I	Site Root	NOL	TL	OthF	TEL	Dcs	Imgs	NOP	NAL	Time/Seconds
1	phuc.qu.edu.sa	1230	211	2	6	1	57	145	205	13.97
2	ucc.qu.edu.sa	1507	266	2	18	1	57	188	248	9.55
3	bhsc.qu.edu.sa	2758	325	4	14	8	86	215	311	22.94
4	dhe.qu.edu.sa	2878	541	2	16	2	275	246	525	40.3
5	cavm.qu.edu.sa	3555	371	2	15	11	59	284	356	32.31
6	adc1.qu.edu.sa	3394	412	2	11	5	98	296	401	41.65
7	uasc.qu.edu.sa	3790	383	2	8	4	83	286	375	34.88
8	asc.qu.edu.sa	4528	425	3	3	3	82	335	422	39.77
9	coe.qu.edu.sa	3539	431	2	5	6	81	337	426	34.11
10	pharmacy.qu.edu.sa	4639	597	2	14	93	133	355	583	31.1
11	dent.qu.edu.sa	5776	533	2	13	69	79	363	520	33.44
12	enuc.qu.edu.sa	5886	466	2	6	2	86	370	460	57.82
13	chr.qu.edu.sa	5546	515	2	7	32	93	381	508	43.29
14	fcohsb.qu.edu.sa	5973	526	4	20	26	87	391	506	41.91
15	qec.qu.edu.sa	5343	539	2	17	42	73	405	522	32.28
16	bcc.qu.edu.sa	8605	627	1	1	29	164	431	626	42.14
17	nursing.qu.edu.sa	9061	533	2	3	11	69	448	530	44.71
18	mduc.qu.edu.sa	9480	542	2	7	0	72	461	535	47.66
19	cos.qu.edu.sa	5346	709	2	4	83	66	554	705	56.17
20	cams1.qu.edu.sa	9635	979	2	12	108	202	655	967	58.07
21	csi.qu.edu.sa	9445	981	3	16	118	78	766	965	72.05
22	coc.qu.edu.sa	10567	979	12	9	5	131	822	970	64.87
23	asoj.qu.edu.sa	2780	383	2	8	3	78	261	344	34.88
24	uasc.qu.edu.sa	2730	425	3	3	3	72	315	292	39.77
–	Average	5567	541	3	10	30	101	397	530	41

B. Feature Extraction

A set of properties organized as logical rules discovered from ODS contents. This set consists of five mathematical rules. These rules represent a special type of advanced knowledge. It was structured based on the internal relations of hyperlinks. The formulas from 1 to 5 show this type of results. Before the extraction of the required features from the ODS contents, the set theory was used to represent the basic components, as follows.

Set of External links: $EL_i := \{EL_1, EL_2, EL_3, \ldots EL_n\}$

Set of Media:=Set of Images+Set of Videos+Set of Audios:=

$\{Im_{i+} V_i + Ao_i\}$, *Where*

$Im_{i:=}\{ Im_1, Im_2, Im_{3,\ldots} Im_n \}$

$V_{i:=}\{ V_1, V_2, V_{3,\ldots} V_n \}$

$Ao_{i:=}\{ Ao_1, Ao_2, Ao_{3,\ldots} Ao_n \}$

Set of Files_Doc, f_i *_D* $:= \{F_1, F_2, F_{3,}\ldots F_n\}$

Set of Other Files, $Othr_i := \{Othr_1, Othr_2, Othr_3, \ldots Othr_n\}$

Set of Leaves (Paths), $Lv_i := \{ Lv_1\ Lv_2\ Lv_3 \ldots Lv_n\}$

The analysis of each component takes a period of time (T), the total analytical time required to analysis a sample of websites is T_i. Where, $T_{i:=}\sum(\{T_1, T_2, T_3, \ldots T_n\})$, also the web structure organized in different levels and different tracks, where, Track Length, $Tr_L := \{Tr_1, Tr_2\ Tr_{3,}\ldots Tr_n\}$.

Five features are extracted, to reflect the internal behavior of websites structure. This in turn constitutes a special type of knowledge organized as a set of relations/rules, concluded based on the contents of ODS and set theory as follows:

N_ of Total_Links=Active_links+External_links (1)

Active_links =Total_No of Pages + No of Other attributes (2)

External_Links= Total_Links – Active_Links (3)

The additional features (AF) are represented in rule 4.

$$(AF) = \sum_{i=1}^{n} Audio_i + \sum_{j=1}^{k} Vedio_j + \sum_{c=1}^{l} Movies_c +$$

$$\sum_{o=1}^{m} Doc\_Fs_o + \sum_{q=1}^{z} Others\_Fs_q + \sum_{r=1}^{i} \mathrm{Im}ages_r \quad (4)$$

No of Pages= Active_links –Other attributes (5)

VIII. Results Discussion and Interpretation

The analyzed attributes belong to a set of 24 homogeneous websites of academic colleges within the Qassim University websites, as presented in Table I. The figures from 3 to 6 show some of the inter-relationships in terms of the attributes that are provided for comparisons and extracted from 24 websites. Figure 3 illustrates the relation between the total links, the external links, and the active links. These details are shown in rule number 3. Figure 4 presents the distribution of the websites' components, such as number of active links, documents, images, number of pages, and other files.

In Figure 4, the values range from largest values to smallest, starting with active links and then ending with other files.

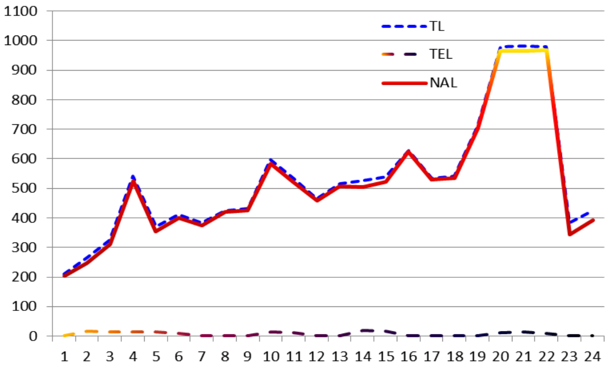


Fig. 3. Rule1 & Rule 3 (total links and external links)

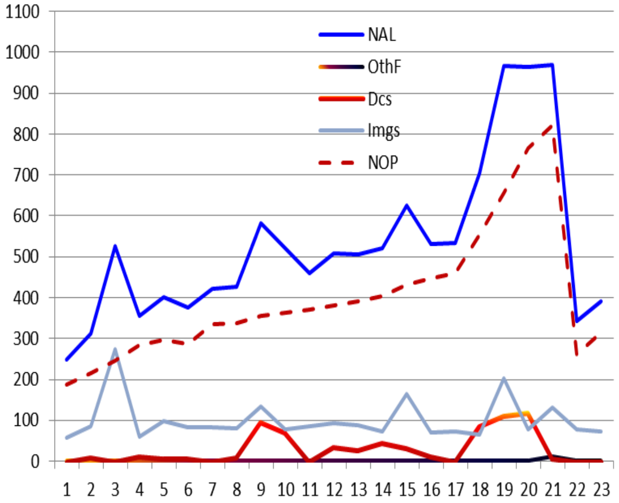


Fig. 4. Rule number 2

As well as Figure 5 represents the rule no 4 and shows the relation between three ingredients of websites structure, such as images, docs, and other files. The attributes "images" have obtained the highest level, while "other files" have obtained the lowest level in this comparison. Figure 6 represents the rule number 5. It shows the relation between the active links, documents, images, and other files.

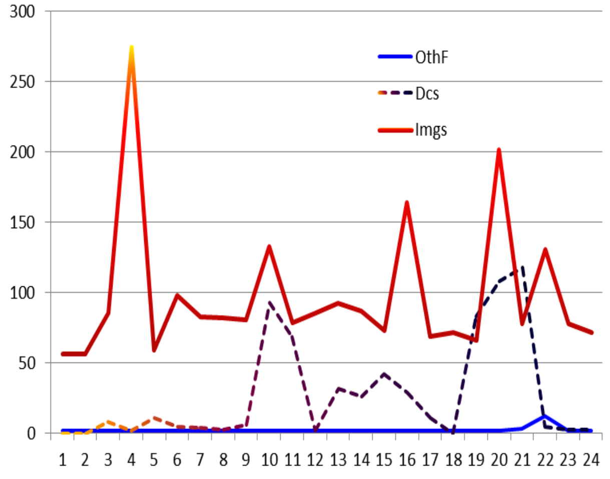


Fig. 5. Rule no.4

It found that, the number of active links is higher than the number of pages in each website structure, because active links encompass many attributes; the number of pages, docs, all types of media files, and other files. Also, it found that, some attribute values increased in some colleges and decreased in others based on the nature of the college. So, the internal relations among the website components are reflecting the development requirements of any website. The relations obtained above in the five formulas are new features that have been deduced based on the contents of ODS.

These formulas represent five association rules. These rules are reflexing the behavior of hyperlinks components and the nature of websites structure.

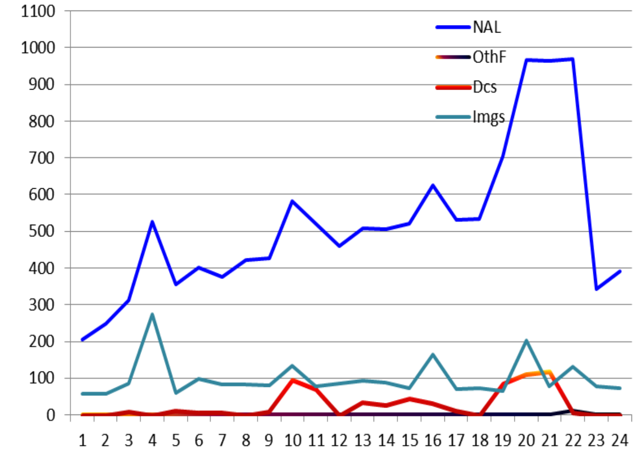


Fig. 6. Representation of Rule 5 components

These formulas can be used later for further analysis and websites estimation, such as structure size estimation.

IX. Conclusions

The research established an algorithm to produce a suitable repository of research data. The proposed algorithm can be applied for any type of websites to analyze the whole hyperlinks and extract many attributes related to the website structure and its contents. About 127991 links were analyzed in this research. It is covered about 69% of the academic colleges of Qassim University. The raw data sets prepared for scientific research purposes. In addition, the research results provide a detailed description of the internal relations of website structure components, where five rules were included in this situation based on the produced ODS. This research has achieved two objectives, based on the analysis of educational websites.

There are many benefits that can be derived from these results, including the ability of developers and users to get a comprehensive understanding of the components of the internal structure of each website and discover the complex relationships between various components. In addition, developers can discover basic relationships that can be employed in the planning and development process of new websites, as well as the results of this research helps developers to build new standards models to estimate different aspects related to the websites developments stages.

X. Future Work

Future research can include the establishment of big ODS based on the proposed algorithm and the implemented tool. These data can be provided as big repositories. These repositories can be analyzed to explore invisible features within different types of websites' structures. The results of this research will help developers in some fields such as websites measurement and estimation models, especially for the early prediction during the development life cycle, for example, these results will be applied in the measurement field, in one of specialized researches which, approved for support by the Deanship of Scientific Research in Qassim University.

References

[1] Anand D., “Improved Collaborative Filtering using Evolutionary Algorithm based Feature Extraction, International Journal of Computer Applications, vol. 64, no.20, 2013.

[2] Benslimane S., Malki M., Rahmouni M., and Rahmoun A., “Towards Ontology Extraction from Data-Intensive web sites: An HTML Forms-Based Reverse Engineering Approach,” The International Arab Journal of Information Technology (IAJIT), vol. 5, no. 1, pp. 34-44, January 2008.

[3] Birla B. and Patel S., “An Implementation on web log mining,” International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, no. 2, pp. 68-73, 2014.

[4] Brin S. and Page L., “The anatomy of a large scale hypersexual web search engine,” Computer Network and ISDN Systems, pp.107-117, 1998.

[5] Chakrabarti S., Dom E., Gibson D., Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., and Tomkins A.,“Mining the link structure of the world wide web,” IEEE Comput., vol.32, pp. 60-67, 1999.

[6] Chopra P. and Ataullah M., “A survey on improving the efficiency of different web structure mining algorithms,” International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no.3, 2013.

[7] Dean J. and Henzinger M., "Finding related pages in the world wide web," Elsevier Science B.V, pp. 389-401, 1999.

[8] Derouiche N., Cautis B., and Abdessalem T., “Automatic extraction of structured web data with domain knowledge,” IEEE 28th International Conference on Data Engineering, 2012.

[9] Getoor L., “Link Mining: A New Data Mining Challenge," SIGKDD Explorations, vol. 4, no. 2, 2003.

[10] Gibson D., Kleinberg J., and Raghavan P., “Inferring web communities from link topology,” Proceeding of the of the 9th ACM Conference on hypertext and hypermedia, June 20-24, ACM Press, PA, USA, pp: 225-234, 1998.

[11] Horowitz E., Sahni S., and Rajasekaran S., “Fundamentals of Computer Algorithms,” Galgotia Publications Pvt. Ltd, pp.112-118, 2008.

[12] Iraklis V., Michalis. V., Maria H., Benjamin N., and Inria F., “A closer view on web content management enhanced with link semantics,” IEEE Trans, 2004.

[13] Jeyalatha S. and Vijayakumar B., ”Design and implementation of a web structure mining algorithm using breadth first search strategy for academic search application,” 6th International Conference on Internet Technology and Secured Transactions, Abu Dhabi, United Arab Emirates, 11-14 December 2011.

[14] Kao H. and Lin S., "A Fast PageRank Convergence Method based on the Cluster Prediction," IEEE/WIC/ACM International Conference on Web Intelligence, IEEE, Computer society, 2007.

[15] Kleinberg J., “Authoritative sources in a hyperlinked environment,” Journal of ACM, vol.46, pp.604-632, 1999.

[16] Kumar R., Raghavan P., Rajagopalan S., and Tomkins T., “Trawling the web for emerging cyber-communities,” IBM Almaden Research Center K53, 650 Harry Road, San Jose, CA 95120, USA, 1999.

[17] Kumar R. and Singh K., “Web structure mining: Exploring hyperlinks and Algorithms for Information Retrieval,” Science Publications, American Journal of Applied Sciences, vol. 7, no. 6, pp.840-845, 2010.

[18] Mishra N., Jaiswal A., and Ambhaikar A., "An effective algorithm for web mining based on topic sensitive link analysis," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 4, April 2012.

[19] Taherizadeh S. and Moghadam N., "Integrating web content mining into web usage mining for finding patterns and predicting users’ Behaviors," International Journal of Information Science and Management, vol. 7, no. 1, January 2009.

[20] Xing W. and Ghorbani A., “Weighted PageRank algorithm,” Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May 19-21, IEEE Computer Society, Washington DC, USA, pp.305-314, 2004.

The Experience-Based Safety Training System Using Vr Technology for Chemical Plant

Atsuko Nakai, Yuta Kaihata, Kazuhiko Suzuki
Center for Safe and Disaster-Resistant Society
Okayama University
Okayama, Japan

***Abstract*—In chemical plants, safety measures are needed in order to minimize the impact of severe accidents and natural disasters. At the same time, carrying out the education and training to workers the corresponding operation in non-stationary situation is essential. However, reproducing the non-stationary conditions to actual equipment or mock-up cannot be performed because it is dangerous. By using the virtual reality (VR) technology, we can build up a virtual chemical plant with lower cost compared to real plant. The operator can experience the fire and explosion accidents in the virtual space. Therefore, in this paper, we propose an experienced-based safety training system for implementing the education and training by using the non-stationary situation in the computer. This proposed system is linked with the dynamic plant simulator. A trainee can learn the correct operation through the simulated experience to prevent an accident. The safety awareness of workers will improve by experiential learning. The proposed system is useful for safety education in chemical plant.**

Keywords—safety education; training system; virtual reality

I. Introduction

The chemical plant has a responsibility to provide and maintain a safe environment for people that live in such circumstances. As well-known many kinds of hazardous materials are under controlled in chemical facilities. Plant safety is provided through inherently safe design and various safeguards, such as instrumented systems, procedures, and training. Safety education/training is essentially important to prevent a severe accident. [1] However, in recent years, the retirement of a skillful operator is advancing in Japan. Because of this, the experts know-how and skill is lost. Skilled operators of superior technology has not been fully passed down to young engineers. Furthermore, immature young operator of the lack of experience is increasing. [2] If an accident occurs in a chemical plant, emergency shutdown is required rapidly. However, as revealed by accident at a nuclear power plant in Fukushima, after the plant has stopped safely, it has likely there resulted in serious damage from the error of the operator. [3] In an emergency, the operator is required to make a quick decision in order to prevent the expansion of the accident. But an untrained operator cannot respond enough to emergencies. Education/training of non-stationary operation for workers is needed. However, reproducing the non-stationary conditions to actual equipment or mock-up cannot be performed because it is dangerous. Also a company requires a huge cost to build a mock-up of the plant facilities for education. This paper will show the training system using virtual reality (VR) technology for chemical plants that we have developed. Figure 1 shows an outline of the training system. This system links the plant model within the VR environment from the process value of a dynamic plant simulator. Order to operate safely the safety device of various in chemical plants, operators have a responsible area of their own. There are two roles can be broadly divided into the plant operator. "Field man" is the role of operating the equipment in the field. Is the role that controls the refineries (control room) in the control center, "Board man". A trainee in the system can experience a cooperative work “Field man ” and “Board man” in the similar environment of the real chemical plant.

This paper shows the safety training system using VR technology for chemical plant. The operator can experience the fire and explosion accidents in the virtual space. The safety awareness of workers will improve by experiential learning. Accidents don't happen the way they're supposed to be according to the manuals. We should consider what to do in case of an emergency.

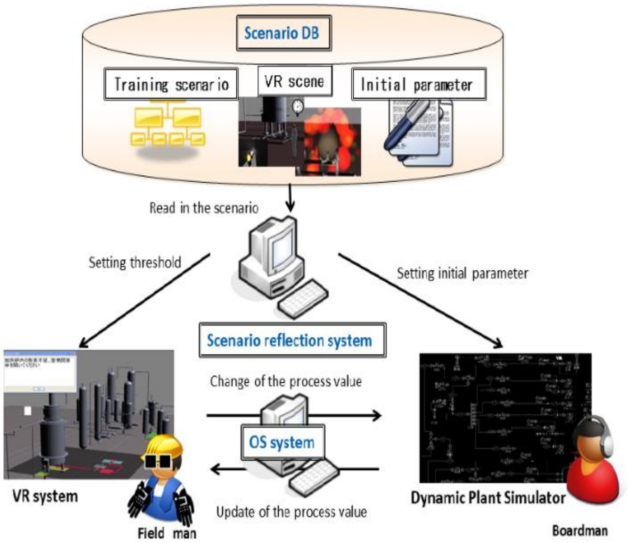


Fig. 1. Outline of the training system

II. Experience-Based Trainig System

A. Outline of the proposed system

The system developed using virtual reality (VR) as new training method. VR is a technology that is artificially creating a sense of reality. [4] This technology is composed of a combination with computer graphics and acoustic efficiency. Thereby, the trainee can get feeling which is actually experiencing things in virtual plant. [5] Also dynamic plant simulator (DS) integrated dynamic simulation environment which reproduces behaviors resembling actual plant operations and provides for a realistic feeling. Therefore, DS has been used for operator training systems. [6][7] This proposed system is composed of a virtual plant that works in concurrence with the dynamic plant simulator(DS) in VR environment. In the result, it became possible to reflect the change of a process value correspond to time progress in the VR environment. [8] DS cannot reproduce severe accidents. Using VR technology, we can express the fire and explosion accidents. Operators/workers can learn the correct operation through the simulated experience to prevent an accident. In this study, the system in consideration of both field man and board man has developed. The board man inputs the process value to the dynamic simulator (DS) display. And the result is reflected in virtual plant. The field man operates the equipment by using the avatar in the Virtual plant through virtual plant screen. An event is qualitatively progress by reading animation in the virtual plant and reading the process value corresponding to the DS according to the amount of operation change. The training system is constructed by three subsystems “scenario database(DB)”, “Operating Support(OS) System” and ” Scenario reflection system”, and two parts, “Virtual plant“ and “Dynamic Simulator ”. The biggest difference between Web learning and VR environment, a trainee can reach out into the virtual world using VR technology. [9]

B. The flow of training

In order to convey training purposes, the system makes a training scenario to provide various information to the operator. Before training start, two or more training scenarios which described the branch criteria by threshold and the event contents are created. And, those are stored in the scenario DB. The flow of training is shown in Figure 2. Next, each VR scene cleated and initial parameters calculated in advance. The plant model in a virtual plant and the process value of DS are linked by OS system. Therefore, the change of the process value in a real plant is reflected to virtual plant. Scenario reflection system compares threshold which writing a select scenario and process value change of DS. When the process values meet the branch criteria, the event is progressing on the qualitatively of update VR scene and an initial parameter to correspond with the scene. In case the process value meets the branch criteria, the VR scene and the parameter of the process value are read by next VR plant and DS parameter. Hereby, it is possible for trainer to focus more to learn important event and to cut down the training time of the plant abnormality.

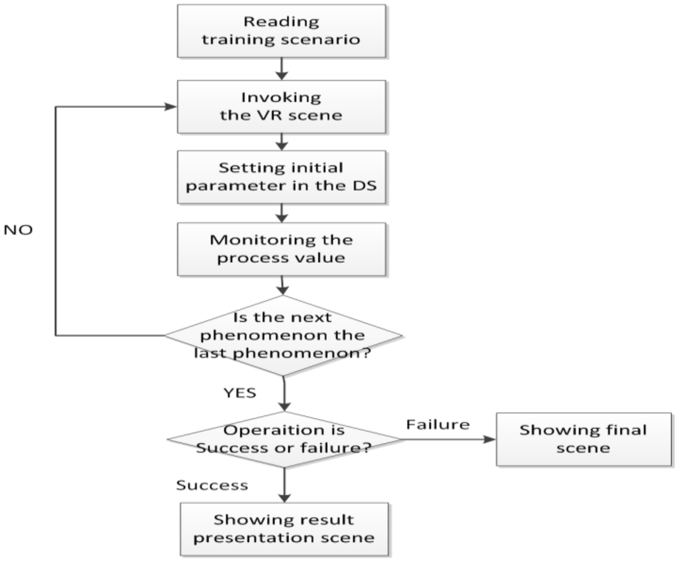

Reading training scenario
Invoking the VR scene
Setting initial parameter in the DS
Monitoring the process value
Is the next phenomenon the last phenomenon?
NO
YES
Operaition is Success or failure?
Failure
Showing final scene
Success
Showing result presentation scene

Fig. 2. The flow of training

C. Scenario database(DB)

This chapter describes about the scenario data base in the training system. “Training scenario”, “VR scene” and “initial parameters” are saved in scenario DB. And those are compatible with branch criteria so as to output each event. A training scenario is written over XML. Branch criteria and plant mode on the basis of flow rate, pressure and temperature, etc. are written in training scenarios. Scenario database has the VR scenes such as plant operation scene, the final scene and result presentation scene. VR scenes are created along the training scenario. Training is started from the plant scene. When the training starts, the display showed a plant scene. Trainer controls an avatar and makes a tour of inspection through the work site. When avatar approaches equipment like a valve in plant scene, equipment operation scene is displayed on the display. This training system updates corresponding scene over and over. This system shows the result presentation scene when a trainee success in early detection and accurate correspondence operation. Result presentation scene is displayed on the assembly operation that trainer went for. Conversely, this system shows a last phenomenon scene when a trainee fails in abnormal detection and correspondence operation. After that, the scene shift to the result presentation scene.

D. Operation support(OS) system

This system reflects the change of the process value into the passage of time in the VR environment by synchronizing the plant model in the VR environment with the process value of the plant simulator. Each process value of the plant simulator is stored in the Excel. The system the outputs process value as a CSV data form. And the VR environment reflects the behavior as a change of the process value of the equipment by storing the value in the array in the VR environment. Moreover, the equipment operation was synchronized with the simulator. When the equipment operation in the VR environment was done, the OS system writes the data in the CSV for the dynamic simulator operation through the Excel.

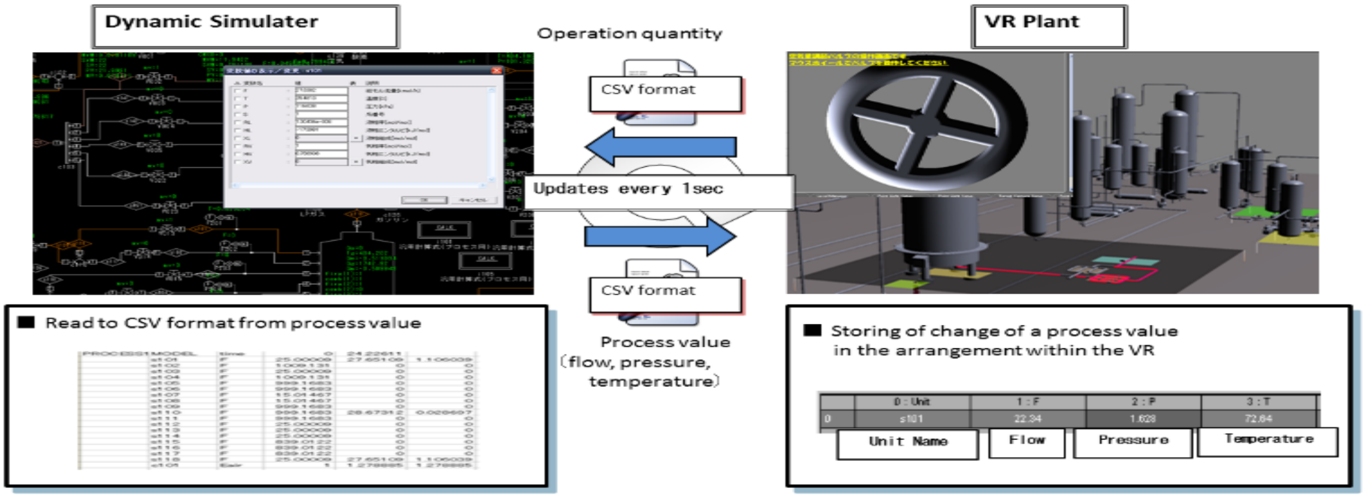


Fig. 3. Composition of OS system

A trainee can operate from the start/stop switch of the operation of valve and pump in the VR environment. Figure 3 shows the composition of OS system.

E. Scenario reflection system

In order to show a trainer an event continuously, it is necessary to output the initial parameter of VR scene and a process value for every event in accordance with the branch condition in a training scenario. The motion of the system from a training start is described along Figure 4. If a process value meets the branch criteria described by the training scenario by a trainer's equipment operation, a simulation will be stopped by suspending DS. Next, VR scene corresponding to the event 2 is read into the VR environment. The initial parameter of the process value corresponding to the event 2 is read into DS, and a simulation is resumed. Thus, the event divided by the threshold of the process value. The event can be continuously shown by making a simulation based on the initial parameter of the process value which differs from each event. [8]

III. Example Of Application

A. The application of the hydrodesulfurization process

A hydrodesulfurization process is targeted in this system.

This process is the reaction with hydrogen separating and removing the sulfur content in a diesel using the metal catalyst of nickel, cobalt, molybdenum, etc. under high temperature and high pressure, and producing products, such as low kerosene of sulfur content, and light oil. The outline of a process is shown in Figure 5.

B. Creation of VR scene based on the training scenario

The training scenario that performs the abnormal detection and a correspondence operation is created. The VR scene express a heating furnace burner for this process. The system used malfunction scenarios from DS for education. [6] When a plant is in a stationary state, the air content in a heating furnace and the flame of the burner are kept normal. As for the flame of the burner, the size is kept at 1/2 from 1/3 to an ignition box ceiling in orange.

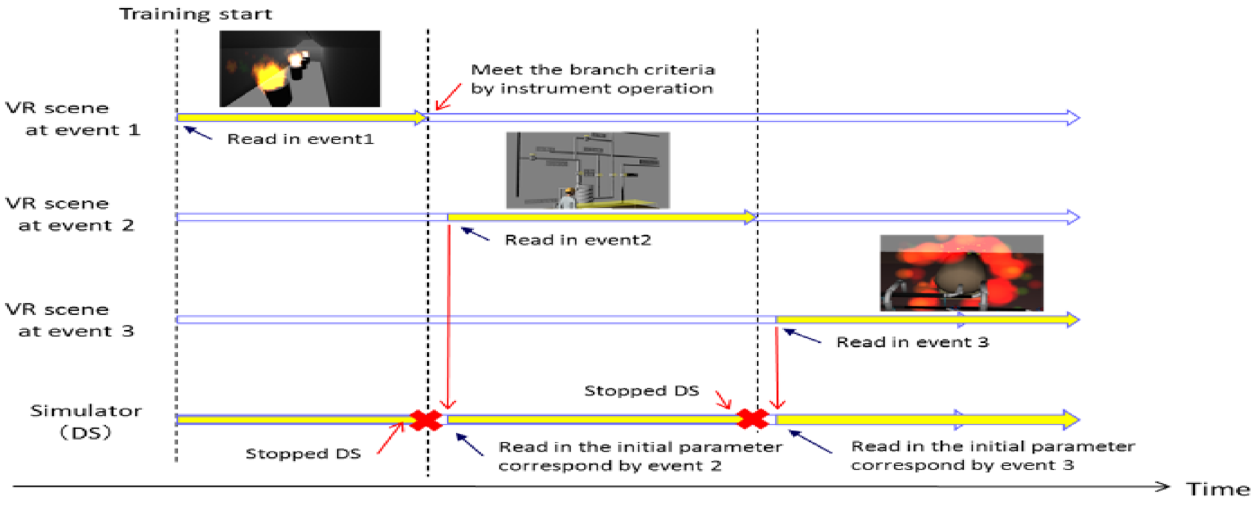


Fig. 4. Operation of the system in alignment with the time

However, when supply of the air to fuel gas decreases, the content air in the heating furnace decreases.

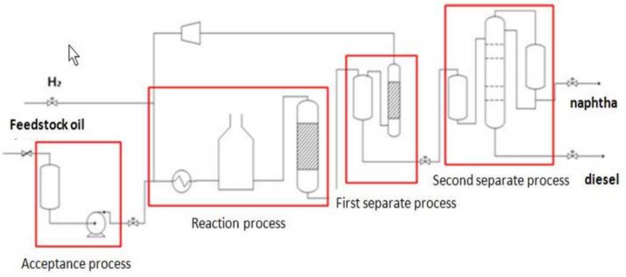


Fig. 5. The outline of a hydrodesulfurization process

Hereupon, the ignition box ceiling changes from a normal orange flame to yellow. Finally a burner is fired by negligence and there is fear of heating furnace explosion. The flame of the burner must be returned to normal color but in this situation it isn’t so. When the fire of the burner goes out, the operator cut off the supply of fuel immediately. Next, in order to remove the fuel gas in the heating furnace, it is necessary to take steam purge 15minutes or more. When the work is not done, combustible gas is full in a heating furnace. And it may be connected to a fire. The necessary scene is created based on the training scenario. The user interface is created so that trainer can check the information on the inside of a heating furnace and the instrumentation equipment in the display. As an accident which occurs in this example, the explosion by gas ignition in the heating furnace is assumed. So, on the last phenomenon scene, VR scene of the explosion, fire of the heating furnace as shown in Figure 6 is created.

C. Integrate Dynamic plant simulator(DS)and VR scene

When a training is carried out, the plant model and process value based on DS is linked by OS system. Accordingly, change of the process value of a real plant is reflected in VR plant. The situation of the link between DS and VR is shown in Figure 7. Pressure, temperature, and concentration are linked between the DS and the VR environment.

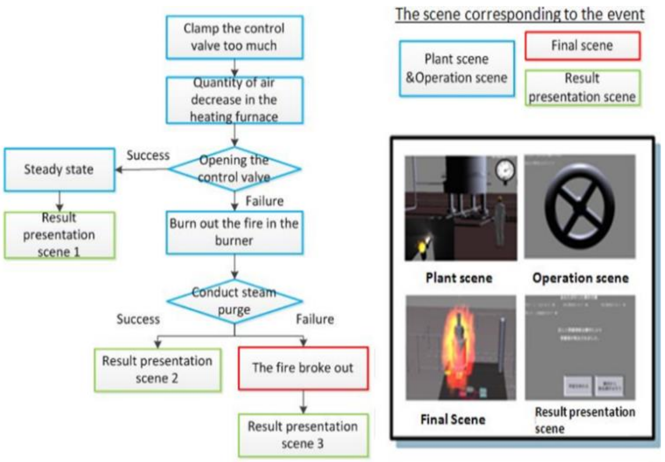

Clamp the control valve too much
Quantity of air decrease in the heating furnace
Steady state
Success
Opening the control valve
Failure
Result presentation scene 1
Burn out the fire in the burner
Conduct steam purge
Success
Failure
Result presentation scene 2
The fire broke out
Result presentation scene 3
The scene corresponding to the event
Plant scene &Operation scene
Final scene
Result presentation scene
Plant scene
Operation scene
Final Scene
Result presentation scene

Fig. 6. Correspondence relation between training scenario and VR scene

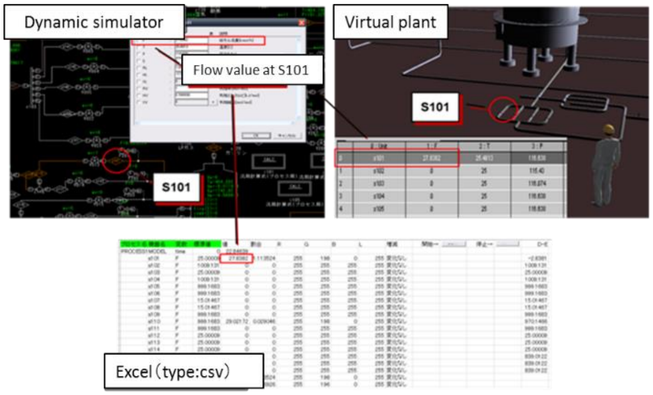
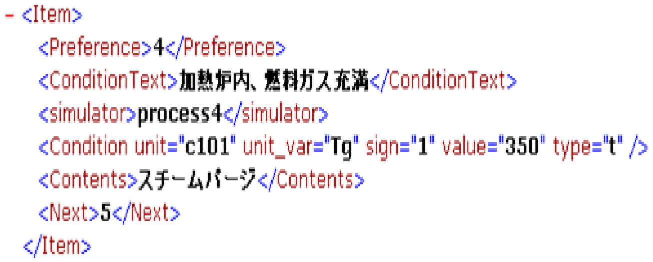


Fig. 7. The link of the process value of the DS and the VR environment

A VR event is developed by comparing the branch condition written in training scenario and the process value reflected in the VR environment. The following Figure 8 describes the event number 4 progress to the event number 5 in the training scenario.



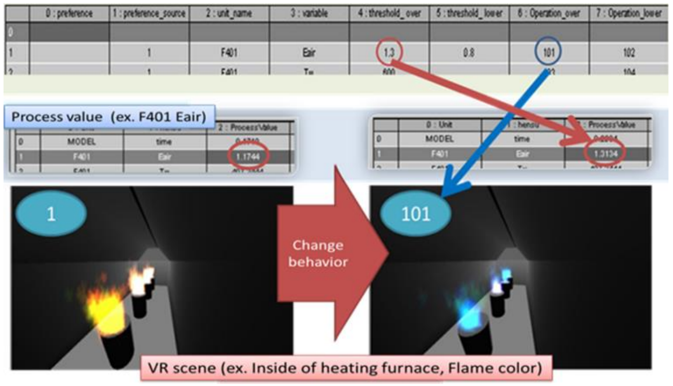
```
- <Item>
    <Preference>4</Preference>
    <ConditionText>加熱炉内、燃料ガス充満</ConditionText>
    <simulator>process4</simulator>
    <Condition unit="c101" unit_var="Tg" sign="1" value="350" type="t" />
    <Contents>スチームパージ</Contents>
    <Next>5</Next>
  </Item>
```

Fig. 8. The contents of the event number 4

If the number 4 event starts, the initial parameter value shows as process4. VR scene corresponding to process4 describes in the <simulator>tag. And VR scene, bases on the <Condition unit="c101"unit_var="Tg"sign="1"value="350"type="t"/>tag in the training scenario. This system supervises the process value Tg of “c101”. “c101” shows a heating scene. If this condition is fulfilled by equipment operation into the VR environment, an event will progress to No. 5 from No. 4.

D. Output of VR scene linked DS

In this system, the process value has linked with the plant model in the VR environment by the above-mentioned OS system between the dynamic plant simulator. The scenario reflection system compares the branch conditions in the scenario by the process value stored in the array in the VR environment and outputs the corresponding scene.



	0 : preference	1 : preference_source	2 : unit_name	3 : variable	4 : threshold_over	5 : threshold_lower	6 : Operation_over	7 : Operation_lower
0								
1		1	F401	Eair	1.3	0.8	101	102
2		1	F401	Tw	600		103	104

Fig. 9. Output of VR scene linked DS

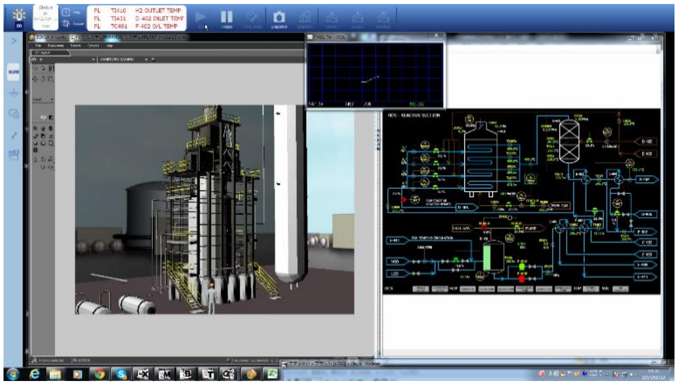


Fig. 10. The training system of the hydrodesulfurization process

Figure 9 describes the system can change the output based on the process value. Figure 10 shows the screen of the actual training system of the hydrodesulfurization process. In Figure 10, VR scene for "Field man" and DS panel for "Board man" are both in the same screen. The developed system can display VR scene and DS panel separate. Boardman and field man can learn operation another role in this training system.

IV. Conclusion

In the chemical plant, rapid generation change of operators is progressing in Japan. The inheritance technology and accomplishments that have been worked in the chemical plant is essential to take advantage in order to extend the competitive production activities. When the accident/disaster happen, the corresponding operation without human error is required in the chemical plant. However, at present, sufficient training is not done for safety and cost issues. In this study, the Experience-based safety training system using VR technology is developed. By using the VR technology, costs and location of the problem can be solved. The trainee can safely perform the training of the corresponding operation against the non-steady situation. In this system, virtual reality environment and dynamic plant simulator are integrated. A trainee in the system can experience a cooperative work "Field man " and "Board man" in a similar environment to the real chemical plant. The scenario reflection system compares the branch conditions in the scenario by the process value stored in the array in the VR environment and outputs the corresponding scene. Virtual plant can express the effects of natural disasters, not only reproduce the accident. Through training using a virtual plant, a trainee can experience the hazardous conditions that cannot be experienced in real. This training is useful for safety education/training for the workers who have no experience. In future this system will be extended to accommodate operation of two people at the same time. After the experience-based learning in virtual plant, the operator should review the corresponding operation to the accident. We will add a re-learning part into this system.

The proposed system will lead to the improvement of the safety awareness of day-to-day operations, consisting of knowledge, skill, sensibility, "total power".

References

[1] S. Jürgen, "Process and Plant Safety – Research & Education Strategy to Keep Long Term Competences", Chemical Engineering Transactions, vol.31, 2013.

[2] T. Ogawa, "The security measure subject of the latest chemical factory", SCAS NEWS 2003 pp.1-2.

[3] Investigation Committee on the Accident at Fukushima Nuclear Power Stations of Tokyo Electric Power Company, Executive Summary of the Final Report, 23 July 2012.

[4] D. Schofield, "Mass Effect: A Chemical Engineering Education Application of Virtual Reality Simulator Technology", MERLOT Journal of Online Learning and Teaching vol. 8, No. 1, March 2012.

[5] A.Wasfy, T. Wasfy , A. Noor, "Intelligent virtual environment for process training", Advances in Engineering Software, No. 35, 2004.

[6] G. Fukano, K. Yokoyama, Y. Yahata, "MIRROR PLANT On-line Plant Simulator and its Applications", Yokogawa Technical Report English Edition ,vol.56 No.1, 2013 pp. 11- 14.

[7] S. Nazir, S. Colombo, D. Manca, "Minimizing the Risk in the Process Industry by Using a Plant Simulator: a Novel Approach", Chemical Engineering Transactions, vol.32, 2013 pp.109-114.

[8] K. Yamamoto, A. Nakai, K. Suzuki, "Development of Experienced-based Training System combined with Process Dynamic Simulation", Asia Pacific Symposium on Safety 2013, October 2013.

[9] C. Norton, I. Cameron, C. Crosthwaite, N. Balliu, M. Tade, D. Shallcross, A. Hoadley, G. Barton, J. Kavanagh, "Development and deployment of an immersive learning environment for enhancing process systems engineering concepts", Education for Chemical Engineers,vol. 3, 2, Decwmber 2008 pp.75-83.

Evaluating Arabic to English Machine Translation

Laith S. Hadla
Department of Translation
Faculty of Arts
Zarqa University
P. O. Box 132222
13132 Zarqa - Jordan

Taghreed M. Hailat
Faculty of IT and CS
Yarmouk University,
Irbid 211-63, Jordan

Mohammed N. Al-Kabi
Faculty of Sciences & IT
Zarqa University
P. O. Box 132222
13132 Zarqa - Jordan

***Abstract*—Online text machine translation systems are widely used throughout the world freely. Most of these systems use statistical machine translation (SMT) that is based on a corpus full with translation examples to learn from them how to translate correctly. Online text machine translation systems differ widely in their effectiveness, and therefore we have to fairly evaluate their effectiveness. Generally the manual (human) evaluation of machine translation (MT) systems is better than the automatic evaluation, but it is not feasible to be used. The distance or similarity of MT candidate output to a set of reference translations are used by many MT evaluation approaches. This study presents a comparison of effectiveness of two free online machine translation systems (Google Translate and Babylon machine translation system) to translate Arabic to English. There are many automatic methods used to evaluate different machine translators, one of these methods; Bilingual Evaluation Understudy (BLEU) method. BLEU is used to evaluate translation quality of two free online machine translation systems under consideration. A corpus consists of more than 1000 Arabic sentences with two reference English translations for each Arabic sentence is used in this study. This corpus of Arabic sentences and their English translations consists of 4169 Arabic words, where the number of unique Arabic words is 2539. This corpus is released online to be used by researchers. These Arabic sentences are distributed among four basic sentence functions (declarative, interrogative, exclamatory, and imperative). The experimental results show that Google machine translation system is better than Babylon machine translation system in terms of precision of translation from Arabic to English.**

Keywords—component; Machine Translation; Arabic-English Corpus; Google Translator; Babylon Translator; BLEU

I. INTRODUCTION

Machine translation means the use of the computers to translate from one natural language into another. Machine translation dated back to the fifties. Although the translation accuracy of online machine translation (MT) systems is lower than translation accuracy of professional translators, these systems are widely used by different people around the world due to their speed and free cost. The translation process in its own is not a straight forward task, for the order of the target words, and the appropriate choice of target words, essentially affect the accuracy of the outputs of the machine translation systems.

Online machine translators rely on different approaches to translate from one natural language into another, these approaches are Rule-based, Direct, Interlingua, Transfer, Statistical, Example-based, Knowledge-based, and Hybrid Machine Translation (MT).

Nowadays, automatic evaluation methods of Machine Translation (MT) systems are used in the development cycle of Machine Translation (MT) systems, system optimization, and system comparison. The automatic evaluation of machine translation systems is based on a comparison of MT outputs and the corresponding professional human translations (Reference translations). Automatic evaluation of machine translation systems offers fast, inexpensive, and objective numerical measurements of translation quality. The first methods to automatic Machine Translation evaluation are based on lexical similarity. These are known as Lexical measures (*n*-gram-based measures) and are based on lexical matching between MT systems outputs and corresponding reference translations [1].

Bilingual Evaluation Understudy (BLEU) is based on string matching, and it is the most widely-used evaluation method to automatically evaluate machine translation systems, and therefore it is used in this study. BLEU is claimed to be language independent and highly correlated with human evaluation, but a number of studies show several pitfalls [1] [2]. BLEU measures the closeness of the candidate output of the machine translation system to reference (professional human) translation of the same text to determine the quality of the machine translation system. The modified *n*-gram precision is the main metric adopted by BLEU to distinguish between good and bad candidate translations, where this metric is based on counting the number of common words in the candidate translation and the reference translation, and then divides the number of common words by the total number of words in the candidate translation. The modified n-gram precision penalizes candidate sentences found shorter than their reference counter parts; also it penalizes candidate sentences which have over generated correct word forms.

Arabic language is a native language of over 300 million people, and it is the most spoken Semitic language. It is the official language of twenty seven countries, and it is one of United Nations (UN) official languages. Moreover, Muslims around the world use it to practice their religion. Modern Standard Arabic (MSA) is used nowadays in Books, Media, Literature, Education, official correspondences, etc. MSA is derived from Classical Arabic (CA). Arabic language is different from English Language, starting with distinctive features of Arabic script: Arabic language alphabets are twenty-eight, Arabic is written from Right to Left as other Semitic languages, Arabic letters within words are connected

This research is funded by the Deanship of Research and Graduate Studies at Zarqa University / Jordan.

in cursive style, short vowels are normally invisible, and finally Arabic language has no uppercase and lowercase letters (no capitalization in Arabic script) [3].

Many studies present different methods to improve machine translation of Arabic into other languages like Carpuat, Marton, and Habash [4], Al Dam, and Guessoum [5], Riesa, Mohit, Knight, Marcu [6], Adly and Al-Ansary [7], and Salem, Hensman, and Nolan [8].

This paper aims to evaluate the effectiveness of two free online machine translation systems (Google Translate (https://translate.google.com) & Babylon (http://translation.babylon.com/)) to translate Arabic to English. The necessary resources to accomplish this study like a dataset of Arabic sentences with two English reference translations are not found. Therefore this study includes a creation of a dataset consisting of 1033 Arabic sentences distributed among four basic sentence functions (declarative, interrogative, exclamatory, and imperative).

This study is organized as follows: section 2 introduces the related work, section 3 presents framework and methodology of this study, section 4 presents the evaluation of two free online machine translation systems under consideration using a system designed and implemented by the second author, section 5 presents the conclusion from this research, and, last but not least, section 6 discusses extensions of the this study and the future plans to improve it.

II. Related Work

Three main categories are used to evaluate machine translation (MT): human evaluation, automatic evaluation, and embedded application evaluation [9]. This section presents a number of related studies to this study that means presenting studies concerned with automatic evaluation of machine translation quality only. Studies related to Bilingual Evaluation Understudy (BLEU) as a method to automatically evaluate machine translation quality are presented first. This section also presents some of the studies related to the automatic evaluation of MT that includes Arabic.

It is usual to have more than one perfect translation of a given source sentence. According to this fact Papineni et al. [2] casted BLEU in 2002 as an automatic metric that uses one or more reference human translation beside a candidate translation of an MT system. The increase in the number of reference translations leads to increase the value of this metric. BLEU metric aims to measure the closeness of a machine-translated (candidate) text to a professional human (reference) translation. BLEU uses a modified precision for *n*-grams at a sentence level and then averages the score over the whole corpus by taking the geometric mean, with *n* from 1 to 4. The BLEU metric ranges from 0 to 1 (or between 1 and 100). BLEU is insensitive to the variations of the order of *n*-grams in reference translations.

There are several studies in the literature presenting enhanced BLEU methods, and in this section only three of these are presented due to the limitation of space.

The first study is conducted by Babych and Hartley [10] and aims to enhance BLEU with statistical weights for lexical items (*tf-idf* and *S*) scores. This enhanced model helps to measure translation adequacy, and uses only one human reference translation, and it is more practical than baseline BLEU metric and more effective. Their enhanced model proposed a linguistic interpretation that relates frequency weights and human intuition about translation Adequacy and Fluency. They used DARPA-94 MT French–English evaluation corpus that has 100 French news texts, where average number of words in each of those French news texts is 350 words. Each French news text is translated by five MT systems, and four of these MT translations are scored by human evaluators. The DARPA-94 MT French–English evaluation corpus has two professional human (reference) translations for each news text. They concluded that their model is consistent with baseline BLEU evaluation results for Fluency and outperform the BLEU scores for Adequacy. They also concluded that their model is reliable if there is only one human reference translation for an evaluated text.

The second study is conducted by Yang, Zhu, Li, Wang, Qi, Li and Daxin [11] and proposed adopting proper weights to different words and *n*-grams into classical BLEU framework. To preserve the language independence in the framework of BLEU, they introduced only the information of the part-of-speech (POS) and *n*-gram length via linear regression model into classical BLEU framework. Experimental results of their study showed that this enhancement yields better accuracy than the original BLEU.

Chen and Kuhn [12] presented in their study a new automatic MT evaluation method called AMBER. This new method AMBER is based on BLEU, but it has new capabilities like incorporating recall, extra penalties, and some text processing variants. The computation of AMBER is based on multiplying Score by Penalty. The modification includes sophisticated formulas to compute Score and Penalty proposed by these two authors. This modified version of BLEU helps to get more accurate results (evaluations) than the results yield by the original IBM BLEU and METEOR v1.0.

Guessoum and Zantout [13] study presented a methodology for evaluating Arabic machine translation. Those authors evaluated lexical coverage, grammatical coverage, semantic correctness and pronoun resolution correctness. Their approach was used to evaluate four English-Arabic commercial Machine Translation systems; namely ATA, Arabtrans, Ajeeb, and Al-Nakel.

The impact of Arabic morphological segmentation on the performance of a broad-coverage English-to-Arabic Statistical machine translation was discussed in the work of Al-Haj and Lavie [14]. In their work, a phrase based statistical machine translation was addressed. Their results showed a difference in BLEU scores between the best and worst morphological segmentation schemes where the proper choice of segmentation has a significant effect on the performance of the SMT.

Professional human translations (Reference translations) are essential to use BLEU method, but not all automatic evaluation MT metrics need reference translations. One of these methods is a user-centered method introduced by Palmer [15]. Palmer's method is based on comparing the outputs of

This research is funded by the Deanship of Research
in Zarqa University / Jordan.

machine translation systems and then ranking them, according to their quality, by expert users who have the necessary needed scientific and linguistic backgrounds to accomplish the ranking process. Palmer's study covers four Arabic-to-English and three Mandarin (simplified Chinese)-to-English machine translation systems.

Most of the people with Arabic as their mother tongue use dialects in their communications at home, markets, etc. One of these dialects is the Iraqi Arabic used mainly in Iraq. To automatically evaluate MT of Iraqi Arabic–English speech translation dialogues, Condon and his colleagues [16] conducted a study and concluded that translation into Iraqi Arabic will correlate higher with human judgments when normalization (light stemming, lexical normalization, and orthographic normalization) is used.

An evaluation of Arabic machine translation based on the Universal Networking Language (UNL) and the Interlingua approach for translation is conducted by Adly and Al-Ansary [7]. The Interlingua approach relies on transforming text in the specified language into a representation form that is language independent that can, later on, be transferred into the target language. Three measures were used for the evaluation process; F_{mean}, F_1, and BLEU. The evaluation was performed using the Encyclopedia of Life Support Systems (EOLSS). The effect of UNL onto translation from/into Arabic language was also studied by Alansary, Nagi, and Adly [17], and Al-Ansary [18].

Carpuat, Marton, and Habash's [4] study is like our study in that it is concerned with translation from Arabic to English. Those authors addressed the challenges raised by the Arabic verb and subject detection and reordering in Statistical Machine Translation. To minimize ambiguities, the authors proposed a reordering of Verb Subject (VS) construction into Subject Verb (SV) construction for alignment only which has led to an improvement in BLEU and TER scores.

A good survey study was conducted by Alqudsi, Omar, and Shaker [19]. In their study, the issue of machine translation of Arabic into other languages was discussed. They presented, through their survey, the challenges and features of Arabic for machine translation. Their study also presents different approaches to machine translation and their possible application for Arabic. The survey concluded by indicating the difficulty of finding a suitable machine translator that could meet human requirements.

Hailat, AL-Kabi, Alsmadi, and Shawakfa [20] conducted a preliminary study to compare the effectiveness of two online Machine Translation (MT) systems (Google Translate and Babylon machine translation systems) to translate English sentences to Arabic. BLUE metric is used in their study to automatically evaluate the MT quality. They conclude that Google Translate is more effective than Babylon machine translation.

The study of Al-Kabi, Hailat, Al-Shawakfa, and Alsmadi [21] is the closest related study to this one, and it is an improvement too [20]. In their study they also use two free online MT systems (Google Translate (https://translate.google.com) & Babylon (http://translation.babylon.com/)) to translate English to Arabic, and a corpus consisting of 100 English sentences, and 300 popular English sayings were used. Al-Kabi et al. [21] study concludes that Google Translate is generally more effective than Babylon. The main differences between this study and our study are the size of corpus used in our study is larger than the corpus they collect, and this study is concerned with translation from Arabic to English not translation from English to Arabic.

The study of Al-Deek, Al-Sukhni, Al-Kabi, and Haidar [22] uses ATEC metric to automatically evaluate the output quality of two Free Online Machine Translation (FOMT) systems (Google Translate and IMTranslator). They concluded in their study that Google Translate is more effective than IMTranslator.

III. THE METHODOLOGY

To evaluate the two online MT systems automatically, first we constructed a corpus consisting of exactly 1033 Arabic sentences with two reference (professional human) translations of each Arabic sentence. The two reference translations were conducted by the first author and Dr. Nibras A. M. Al-Omar from Zarqa University. The size of our corpus is 4169 Arabic words, and the number of Arabic unique words is 2539.Table 1 shows the distribution of the Arabic sentences of the constructed corpus among four basic sentence functions. This corpus is uploaded to Google drive server in order to make it accessible to everyone wish to use it. Those who are interested in this corpus can download it using the following URL:

https://docs.google.com/spreadsheets/d/1bqknBcdQ7cXOKtYLhVP7YHbvrlyJlsQggL60pnLpZfA/edit?usp=sharing

TABLE I. DISTRIBUTION OF SENTENCES AMONG FOUR BASIC SENTENCE FUNCTIONS

Basic Arabic Sentence Functions	No. of Sentences
Declarative	250
Interrogative	231
Exclamatory	252
Imperative	250
Total	1033

The construction of the above corpus is followed by accomplishing the following main steps shown in Figure 1. BLEU method is used in this study to automatically evaluate the effectiveness of translation from Arabic to English by the two online Machine Translation (MT) systems (Google Translate and Babylon machine translation systems).

The main steps followed to accomplish this study are presented in Figure 1. First the source Arabic sentence is translated using (Google Translate) and (Babylon machine translation systems) and two professional human translations (Reference translations). Then, these Arabic sentences are preprocessed by dividing the text into different *n*-gram sizes, as follows: unigrams, bigrams, trigrams, and tetra-grams.

After that, the precision for Babylon machine translation system and Google machine translation system were computed for each of the four gram sizes. In the final step, for each of the four *n*-gram sizes, we compute a unified precision score for that size. These values are then compared to decide which of them gets the best translation.

In order to compute the precision score for each of the four *n*-gram sizes, we have to count first the number of common words in every candidate and reference sentence, and then we have to divide this sum over the total number of *n*-grams in the candidate sentence.

To combine the previous precision values in a single overall score (called BLEU-score), we start by computing the Brevity Penalty (BP) by choosing the effective reference (i.e. the reference that has more common *n*-grams) length which is denoted by *r*. Then we compute the total length of the candidate translation denoted by *c*. Now we need to select Brevity Penalty to be a reduced exponential in (*r* / *c*) as shown in equation 1 [2]:

$$BP = \begin{cases} 1 & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c \leq r \end{cases} \qquad (1)$$

The computation of the final BLEU score is shown in formula (2) and it is based on Brevity Penalty (*BP*) shown in formula (1).

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (2)$$

Where $N = 4$ and uniform weights $w_n = (1/N)$ [2].

This indicates that higher BLEU score for any machine translator means that it’s better than its counterparts with lower BLEU scores.

Papineni, Roukos, Ward, and Zhu [2] study noted that the BLEU metric values range from 0 to 1, where the translation that has a score of 1 is identical to a professional human translation (Reference translation) [2].

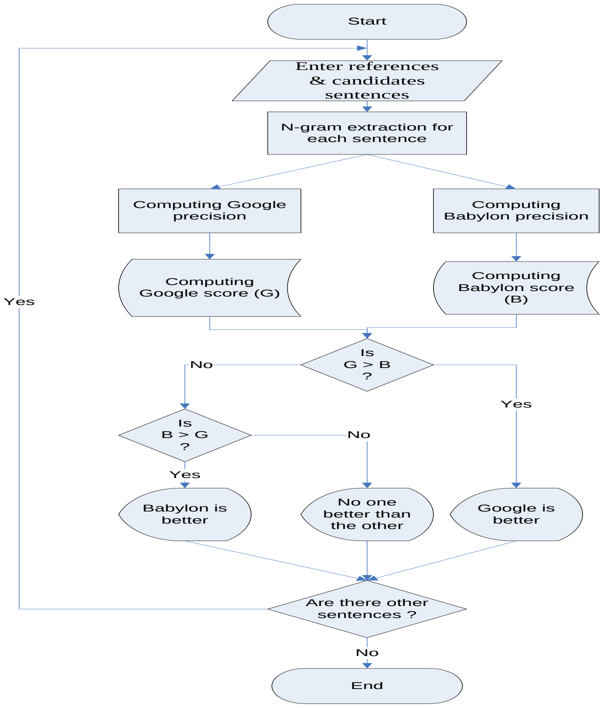

Start
Enter references & candidates sentences
N-gram extraction for each sentence
Computing Google precision
Computing Babylon precision
Computing Google score (G)
Computing Babylon score (B)
Is G > B ?
No
Yes
Is B > G ?
No
Yes
Babylon is better
No one better than the other
Google is better
Are there other sentences ?
Yes
No
End

Fig. 1. Evaluation Methodology Flowchart

IV. The Evaluation

Many automatic evaluation methods of MT are proposed and used during the last few years, beside manual evaluation methods of MT. Bilingual Evaluation Understudy (BLEU) method is one of the well-known automatic evaluation methods of machine translation adopted in this study.

The following notes resulted from the conducted experiments on Google Translate and Babylon machine translation systems:

1) We noticed Babylon MT translates an Arabic word correctly to English, while Babylon MT ignores completely translating those Arabic words in other sentences.

2) Babylon MT could not translate the words that contain related pronouns “الضمائر المتصله”, for example the source sentence in Arabic:” سمعت كلمة اضحكتني”, that was translated using Babylon as: “I heard the word اضحكتنى”

3) Babylon machine translation system could not translate multiple Arabic sentences at one time, while Google Translate can translate a set of Arabic sentences at one time.

In our evaluation and testing of the two MT systems, we found that the translation precision is equal for both MT systems (Google and Babylon) for some sentences, but translation precision of Google Translate is generally better than translation precision of Babylon MT system (0.45 for Google and 0.40 for Babylon).

As a whole, the average precision values of Google and Babylon machine translation system for each type of sentences in the corpus are shown in Table 2. It is obvious that Google Translate system is generally better than Babylon machine translation system, but, as shown in table 2, Babylon MT system is more effective than Google Translate in translating Arabic exclamation sentences into English.

TABLE II. AVERAGE PRECISION FOR EACH TYPE OF SENTENCES

Type Translator	Declarative Sentence	Exclamation Sentence	Imperative Sentence	Interrogative Sentence	Average
Babylon MT System	0.3475	0.3686	0.5189	0.3588	0.39845
Google Translate System	0.4486	0.3378	0.5453	0.4668	0.449625

V. CONCLUSION

Arabic-to-English and English-to-Arabic MT have been a challenging research issue for many of the researchers in the field of Arabic Natural Language Processing (NLP).

In this study, we have evaluated the effectiveness of two automatic machine translators (Google Translate System and Babylon machine translation system) that could be used for Arabic-to-English translation and vice versa.

The accuracy of any MT system is usually evaluated by comparing its outputs to that of professional human translators, or professional human translators can manually evaluate the quality of translation. There is no standard Arabic-English corpus that can be used for such evaluations, therefore, we constructed a corpus and released it for free on the Internet to be used by the researchers in this field.

Although the collected data was relatively small in size, the well-known Arabic sayings usually presented a challenge for the machine translation system to translate them to English, and this problem faces us with these MT systems used to translate English sayings to Arabic.

Although the collected data was relatively small in size, the well-known Arabic sayings usually presented a challenge for the machine translation system to translate them to English, and this problem faces us with these MT systems used to translate English sayings to Arabic.

VI. FUTURE WORK

We plan in the future to study the effectiveness of other automatic evaluation MT methods like METEOR, ROUGE, NIST and RED.

We have tested our experiments on a relatively small corpus, and as part of the future work we are planning to build a larger corpus and release it to be used freely by different researchers in this field.

REFERENCES

[1] J. Giménez, and L. Màrquez, “Linguistic measures for automatic machine translation evaluation,” Machine Translation vol. 24, no. 3-4, pp. 209-240, 2010.

[2] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Stroudsburg, PA, USA, pp. 311-318, 2002.

[3] K. C. Ryding, “A Reference Grammar of Modern Standard Arabic,” Cambridge: Cambridge University Press, 2005.

[4] M. Carpuat, Y. Marton, and N. Habash, “Improving Arabic-to-English Statistical Machine Translation by Reordering Post-verbal Subjects for Alignment,” in Proceedings of the ACL 2010 Conference Short Papers, pp. 178–183, Uppsala, Sweden, 2010.

[5] R. Al Dam, and A. Guessoum, “Building a neural network-based English-to-Arabic transfer module from an unrestricted domain," In Proceedings of IEEE International Conference on Machine and Web Intelligence (ICMWI), pp.94-101, 2010.

[6] J. Riesa, B. Mohit, K. Knight, and D. Marcu, “Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources", In the Proceedings of INTERSPEECH, Pittsburgh, USA, 2006.

[7] Adly, N. and Alansary, S. 2009, “Evaluation of Arabic Machine Translation System based on the Universal Networking Language,” in Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems “NLDB 2009”, pp. 243-257, 2009.

[8] Y. Salem, A. Hensman, and B. Nolan, “Towards Arabic to English Machine Translation,” ITB Journal, Issue 17, pp. 20-31, 2008.

[9] K. Kirchhoff, D. Capurro, and A. M. Turner, “A conjoint analysis framework for evaluating user preferences in machine translation,” Machine Translation, vol. 28, no. 1, pp. 1-17, 2014.

[10] B. Babych, and A. Hartley, “Extending the BLEU MT evaluation method with frequency weightings,” In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04). Association for Computational Linguistics, Stroudsburg, PA, USA, Article 621, 2004.

[11] M. Yang, J. Zhu, J. Li., L. Wang, H. Qi, S. Li., and L. Daxin, “Extending BLEU Evaluation Method with Linguistic Weight,” 2008. ICYCS 2008. The 9th International Conference for Young Computer Scientists, pp. 1683-1688, 2008.

[12] B. Chen, and R. Kuhn, “AMBER : A modified BLEU, enhanced ranking metric,” in Proceedings of the 6th Workshop on Statistical Machine Translation, Edinburgh, UK, pp. 71-77, 2011.

[13] A. Guessoum, and R. Zantout, “A Methodology for Evaluating Arabic Machine Translation Systems,” Machine Translation, issue 18, pp. 299-335, 2005.

[14] H. Al-Haj, and A. Lavie A., “The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation,” vol. 26, no. 1-2, pp. 3-24, 2012.

[15] D. D. Palmer, “User-centered evaluation for machine translation of spoken language," in Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. v/1013- v/1016, 2005.

[16] S. Condon, M. Arehart, D. Parvaz, G. Sanders, C. Doran, and J. Aberdeen, "Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics", Machine Translation, vol. 26, Nos. 1-2, pp. 159-176, 2012.

[17] S. Alansary, M. Nagi, and N. Adly, "The Universal Networking Language in Action in English-Arabic Machine Translation," In Proceedings of 9th Egyptian Society of Language Engineering Conference on Language Engineering, (ESOLEC 2009), Cairo, Egypt, 2009.

[18] S. Alansary, "Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas", in Proceedings of the 11th International Conference on Language Engineering, Cairo, Egypt, 2011.

[19] A. Alqudsi, N. Omar, and K. Shaker, "Arabic Machine Translation: a Survey", Artificial Intelligence Review, pp. 1-24, 2012.

[20] T. Hailat,, M. N. AL-Kabi, I. M. Alsmadi, E. Shawakfa, "Evaluating English To Arabic Machine Translators," 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT 2013) - IT Applications & Systems, Amman, Jordan, 2013.

[21] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating English to Arabic Machine Translation Using BLEU," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 1, pp. 66-73, 2013.

[22] H. Al-Deek, E. Al-Sukhni, M. Al-Kabi, M. Haidar, "Automatic Evaluation for Google Translate and IMTranslator Translators: An Empirical English-Arabic Translation," The 4th International Conference on Information and Communication Systems (ICICS 2013). ACM, Irbid, Jordan, 2013.

AUTHORS PROFILE

Laith Salman Hassan Hadla, born in Baghdad/Iraq in 1970. He obtained his PhD in Machine Translation from Al-Mustansiriya University in (2006), his masters' degree was in stylistic translation from Al-Mustansiriya University in (1995), and his bachelor degree is in Translation from Al-Mustansiriya University in (1992). Laith S. Hadla is an assistant professor at the Faculty of Arts, at Zarqa University. Before joining Zarqa University, he worked since 1993 in many Iraqi and Arab universities. The main research areas of interest for Laith S. Hadla are machine translation, translation in general, and linguistics. His teaching interests fall into translation and linguistics

Taghreed M. Hailat, born in Irbid/jordan in 1986. She obtained her MSc. degree in Computer Science from Yarmouk University (2012), and her bachelor degree in Computer Science from Yarmouk University (2008). Currently, she is working at Irbid chamber of Commerce as a Computer Administrator and previously as a trainer of many computer courses at Irbid Training Center

Mohammed Al-Kabi Mohammed Al-Kabi, born in Baghdad/Iraq in 1959. He obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his masters degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq(1981). Mohammed Naji AL-Kabi is an assistant Professor in the Faculty of Sciences and IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a lecturer in PSUT and Jordan University of Science and Technology (JUST). AL-Kabi's research interests include Information Retrieval, Sentiment analysis and Opinion Mining, Web search engines, Machine Translation, Data Mining, & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Big Data, Web programming, data mining, DBMS (ORACLE & MS Access).

Routing in Wireless Sensor Networks based on Generalized Data Stack Programming Model

Hala Elhadidy1, Rawya Rizk[1]
1Electrical Engineering Department, Port Said
University Port Said, Egypt

Hassan T. Dorrah[2]
[2]Electrical Engineering Department, Cairo University,
Giza, Egypt

***Abstract*—Generalized Data Stack Programming (GDSP) model describes that any affected activity or varying environment is intelligently self-recorded inside the system in form of stack-based layering types where stack is re-defined to be one of six classes. The multi-stacking network is presented thinking of that any system connects to other system to work properly. This addresses a novel way to investigate and analyze any system. Wireless Sensor Networks (WSNs) monitor the environment and take action accordingly. However, WSN suffers from some weakness due to nodes failure or interference which affects the network topology and the routing table at each node. In this paper, the GDSP model is applied on the routing problems in WSNs. A history matrix at the user side is proposed to retrieve and backward the events affected the network.**

Keywords— GDSP; history matrix; multi-stacking network; network topology; routing table; WSN

I. Introduction

Any physical system is a collection of interconnected physical parts or elements to perform desired function [1]. Since all physical elements have properties that change with environment and age, we cannot always consider the parameters of a control system to be completely stationary over the entire operating life of the system [2]. Therefore, the controller in any system updates when new measurement data arrives or at a specific time or both as the system can be event-driven or time-driven system [3].

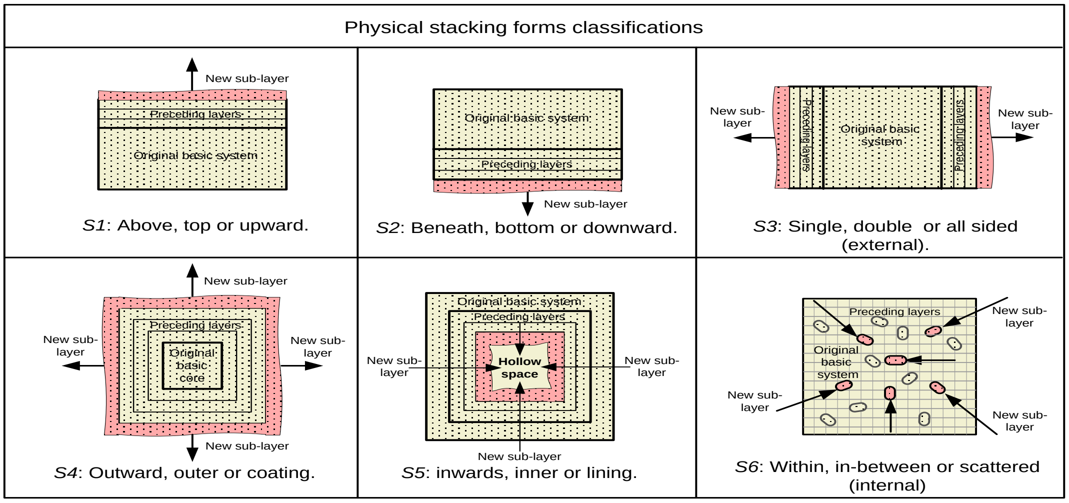
Wireless sensor network (WSN) is an example of the event-driven systems where many of its nodes have the ability to sense data from environment, perform simple computations and transmit this data using wireless interfaces to the base station (sink) either directly or in a multi-hop fashion through neighbors to take action accordingly [4]. All the nodes are working co-operatively to accomplish pre-specified mission objectives, such as minimum exploration time or complete monitoring coverage [5, 6]. Routing in WSNs is a challenge issue [7]. Routing protocols need to adapt to any changes to the application requirements and network dynamics during run-time of WSNs. Routing typically begins with neighbor discovery. Nodes send rounds of messages (packets) and build local neighbor tables. Routing tables in WSNs include the minimum information of each neighbor location, nodes' remaining energy, delay via that node, and an estimate of link quality [8]. Once the tables exist, in most WSN routing algorithms, messages are directed from a source location to a sink address based on geographic coordinates. Due to node mobility and failure in WSN, the network topology may be continuously changing with time and therefore the routing table of each node is also changing dynamically.

On the other hand, system changes are transferred into stack-based changes in physical system layers with different stack classes and categories in time-driven systems. However, system changes are transferred at each event step to corresponding infinite stack-based parameters changes at basic system level and to changes in consolidity index at the higher level. The definition of the stack has been changed to be more realistic and express any system [9, 10]. Based on this new stack definition, a new Generalized Data Stack Programming (GDSP) model was presented in [11]. Since any system has many elements and each of them somehow can be represented by stack, if any operation is done to one of them, it will affect the other neighbors. Therefore the new vision of the system as a network of stacks rises and is covered in this paper. Also, this paper presents how to apply the new theory of the stack on WSNs in the case of changing the topology of the network and how the stacking network works in this case.

The rest of this paper is organized as follows: Section II presents the GDSP model. Section III presents the node failure problem in WSNs. Section IV introduces how GDSP is applicable to WSNs. Section V presents the multi-stack layering network and how it is applied to WSNs. Section VI presents conclusions and future works..

II. Generalized Data Stack Programming (Gdsp) Model

The proposed GDSP model depends on the idea of simulating what happen in the physical world where any object can be a stack. System changes in real life may be logically conceived to be internally stacked in the form of a new sub-layer arranged in some form in relations to the other preceding existing layer(s). This stack uses matrices to represent the system which can grow or shrink according to the change occurred to the system. The matrix formulation allows fast, direct design and reconfiguration of discrete event controllers [12]. It provides a better dynamical description and high-level interface than other popular tools for discrete event systems, such as Petri Nets. The GDSP model includes six classes of stacking which effectively forms *modular blocks* for handling various applications. Fig. 1 shows that the stack is redefined to be one of six classes. At S1, the insertion or deletion can be done at the top. The insertion or deletion can be done also at the bottom, at one or both sides, as an outer ring around the object, as an inner ring inside the object, or at any place in the object at S2, S3, S4, S5, or S6 respectively.



S1: Above, top or upward.

S2: Beneath, bottom or downward.

S3: Single, double or all sided (external).

S4: Outward, outer or coating.

S5: inwards, inner or lining.

S6: Within, in-between or scattered (internal)

Fig.1. Schematic sketches of various physical stacked-based change classifications at the system basic level based on the directions of system sub-layers or sub-stratum change

TABLE.I. GDSP MODEL

Class	Features	Matrix representation	Applications
S1	Traditional form, $S1^{+}$ or $S1^{-}$ done at top of an object.	$\xrightarrow{S1^{+}}$ Growing matrix $\begin{bmatrix} + & + & + \\ x & x & x \\ x & x & x \end{bmatrix}$ $\xrightarrow{S1^{-}}$ Shrinking matrix $\begin{bmatrix} x & x & x \end{bmatrix}$	The computer memory.
S2	$S2^{+}$ or $S2^{-}$ done at the bottom of an object	$[x$ $\xrightarrow{S2^{+}}$ Growing matrix $\begin{bmatrix} x & x & x \\ x & x & x \\ + & + & + \end{bmatrix}$ $\xrightarrow{S2^{-}}$ Shrinking matrix $\begin{bmatrix} x & x & x \end{bmatrix}$	The marine algae under water.
S3	$S3^{+}$ or $S3^{-}$ done to a side or two sides of an object	$\begin{bmatrix} x & x & x \\ x & x & x \end{bmatrix} \xrightarrow{S3^{+}}$ Growing matrix $\begin{bmatrix} + & x & x & x \\ + & x & x & x \end{bmatrix}$ or $\begin{bmatrix} + & x & x & x & + \\ + & x & x & x & + \end{bmatrix}$ $\xrightarrow{S3^{-}}$ Shrinking matrix $\begin{bmatrix} x & x \\ x & x \end{bmatrix}$ or $\begin{bmatrix} x \\ x \end{bmatrix}$	External rust on an industrial pipe.
S4	$S4^{+}$ or $S4^{-}$ done to the outer ring of an object	$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix} \xrightarrow{S4^{+}}$ Growing matrix $\begin{bmatrix} + & + & + & + & + \\ + & x & x & x & + \\ + & x & x & x & + \\ + & x & x & x & + \\ + & + & + & + & + \end{bmatrix}$ $\xrightarrow{S4^{-}}$ Shrinking matrix $[\,x\,]$	The tree trunk.
S5	$S5^{+}$ or $S5^{-}$ done to the inner ring of an object under a condition that the object has an empty center.	$\begin{bmatrix} x & x & x & x & x \\ x & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & x \\ x & x & x & x & x \end{bmatrix} \underset{S5^{-}}{\overset{S5^{+}}{\rightleftarrows}} \begin{bmatrix} x & x & x & x & x \\ x & + & + & + & x \\ x & + & 0 & + & x \\ x & + & + & + & x \\ x & x & x & x & x \end{bmatrix}$	The fats can deposit inside the artery
S6	Add an element at (a,b) then rearrange the matrix horizontally or vertically. Removing an element does not mean to remove a column or row.	$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & 0 \end{bmatrix} \xrightarrow{S6+}$ Horizontall $\begin{bmatrix} x & x & x & 0 \\ x & y & x & x \\ x & x & 0 & 0 \end{bmatrix}$ or vertical $\begin{bmatrix} x & x & x \\ x & y & x \\ x & x & 0 \\ 0 & x & 0 \end{bmatrix}$ $\xrightarrow{S6^{-}}$ Horizontal $\begin{bmatrix} x & x & x \\ x & x & 0 \\ x & x & 0 \end{bmatrix}$ or vertical $\begin{bmatrix} x & x & x \\ x & x & x \\ x & 0 & 0 \end{bmatrix}$	Deposit of sand or minerals in the kidney, fruit worm and having a tumor under skin.

Table 1 summarizes the basic features and the matrix representation of these six classes with their applications. The codes and the mathematical representation for each class are found in [11]. It is assumed that the new layer is uniform – adding a whole row, column or ring depends on the stack class. However, the layer could be non-uniform containing segments of different nature and appearance. This case is exactly like the deposits of fats inside the human artery where the fats are not accumulated in a complete ring shape. So its matrix can be as follow:

$$\begin{bmatrix} x & x & x & x & x \\ x & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & x \\ x & x & x & x & x \end{bmatrix} \text{ will be } \begin{bmatrix} x & x & x & x & x \\ x & x & 0 & 0 & x \\ x & x & 0 & 0 & x \\ x & x & 0 & 0 & x \\ x & x & x & x & x \end{bmatrix} \text{ or } \begin{bmatrix} x & x & x & x & x \\ x & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix} \text{ or even } \begin{bmatrix} x & x & x & x & x \\ x & 0 & x & x & x \\ x & 0 & 0 & 0 & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix}.$$

There are many applications that use more than one stack class like in the soil. The deposits of soil will be added up increasing the height of it ($S1^+$). The decayed animals and plant remains in the soil affected its fertility ($S2^+$). Also there are some insects and animals live inside the soil. The presence of these insects looks like adding an element in a specific place ($S6^+$). WSN is an example of the applications that uses more than one stack class as will be explained in the next section.

If a node failure is discovered in a WSN, one of the solutions to maintain the reliability of the network is that, the nodes following the failed one which are lying in the opposite side of the sink move to new locations to optimize network performance while simultaneously confirming the connectivity requirements. The optimal configuration of the nodes is lying evenly on the line from the source to the sink as proven in [13]. However, moving the nodes could affect the coverage area of the network which has an arbitrary shape and therefore cause degradation in the network performance [14] as shown in Fig. 2.

III. The Node Failure Problem in WSNs

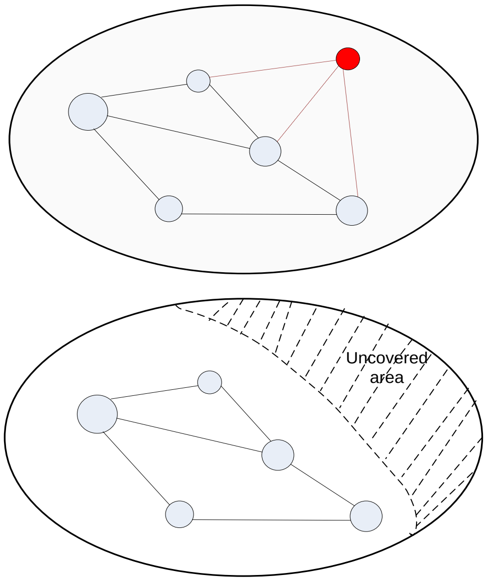


Fig.2. the difference between the covered area of a WSN before and after removing a failed node

The network topology can be represented by a matrix ($N\times N$) where N is the number of nodes. m(a,b)= 0 or 1 where 1 means there is a connection (route) between node a and b. For example, the topology of the network shown in Fig. 3 can be represented as

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

The routing table of each node indicates the minimum information about the neighbors of the node. It may contain the remaining energy of each node as shown in Table 2 that represents the routing table of Node 3.

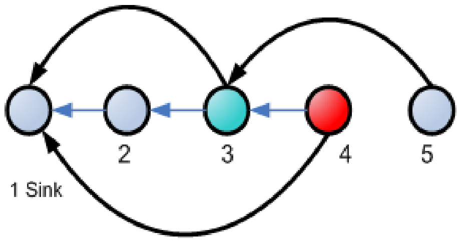


Fig.3. An example of a WSN

TABLE.II. ROUTING TABLE OF NODE 3

Node ID	Position x	Position y	Remaining energy
1	50	50	100
2	60	50	80
4	80	50	10
5	90	50	40

If a failure happens to a node, a message will pass through the network (which happens regularly) to change the topology according to the routing algorithm that is used in this network. In the example shown, if Node 4 has a failure, there are two cases [15]. First, Node 5 will take the position of Node 4 if it is able to move. Then, the topology changes according to its communication range. If this moving node can covers the communication range, then the topology of a WSN becomes as shown in Fig. 4(a) and its matrix can be represented as $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$.

If the sensing area is shrunk which affects the coverage area and then causes degradation in network performance, a new node needs to be inserted at the end of the network as shown in Fig. 4(b) and its final matrix can be considered as $\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$.

Table 3 shows the final routing table of Node 3 in this case. It is shown from the table that, the remaining energy of Node 1 is 100% since it is important to plug in the sink node to be always alive [16] while the remaining energy of the other nodes decreases with the usages.

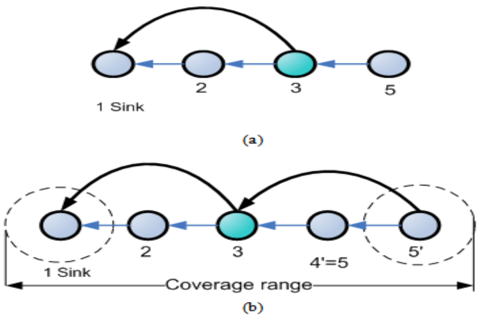


Fig.4. The network after changing the topology if Node 5 is able to move. (a) If the sensing area doesn't change, (b) If a new node is added for complete coverage.

TABLE.III. THE NEW ROUTING TABLE AFTER ADDING A NEW NODE AT THE END

Node ID	Position x	Position y	Remaining energy
1	50	50	100
2	60	50	77
4'(5)	80	50	39
5'	90	50	100

Second, if Node 5 is a static node, a new node needs to be inserted instead of Node 4 and according to its communication range the topology will change as in Figure 5 for example and its matrix can be represented as $\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$.

In this case, the final routing table for Node 3 is shown in Table 4.

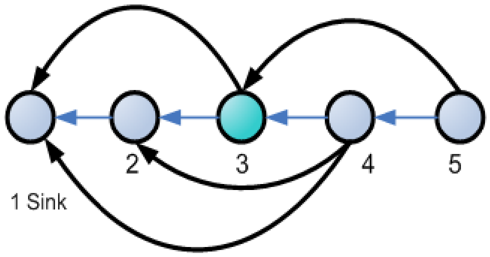


Fig.5. The network after adding a new powerful node at 4 if Node 5 is static

TABLE.IV. THE NEW ROUTING TABLE AFTER ADDING A NEW NODE AT THE POSITION OF NODE 4

Node ID	Position x	Position y	Remaining energy
1	50	50	100
2	60	50	77
4	80	50	100
5	90	50	39

IV. GDSP IMPLEMENTATION ON WSNS

By comparing the previous matrices, we can consider that the WSN is a multi-stacking application of GDSP as any physical system. In the first case, $S6^-$ will be used repeatedly for all nodes to remove the failed one m(i,4) where i=1,2,3,5 and then rearrange horizontally and remove m(4,i) then rearrange vertically. Adding a column at the end ($S3^+$ one side) and row at the end ($S2^+$) with values 0 and 1 according to the new node's sensing rang are equivalent to inserting the new node. In the second case, $S6^-$ will be used repeatedly for all nodes to remove the failed one m(i,4) where i=1,2,3,5 and then rearrange horizontally and remove m(4,i) then rearrange vertically. After that a new node is inserted instead of the failed Node 4 (as in the example) ($S6^+$) so m(i,4) is added with a value 1 or 0 according to the sensing range of the node and in this case will be 1 all the time then rearrange horizontally and add m(4,i)=1 and rearrange vertically.

On the other hand, it is mentioned that if the topology is changed, a round of messages is moved through the network so that each node updates its routing table. To save energy, this message will move around the failed node's neighbors only and their routing tables only are updated. In the first case and when Node 4 is removed, all the following nodes to it will move towards it then add a new one at the end. This means $S6^-$ will be applied to the routing table then rearrange the table vertically with changing the nodes ID then $S2^+$ with the information of the new node.

In the second case, another node (new 4) will be inserted. Therefore, a new entry in the routing table with the same ID of the failed node but with the new information is inserted. Applying GDSP, $S6^-$ is used to remove the failed node without rearranging either horizontally or vertically leaving an empty row. After adding a new node, its information will be in this empty row ($S6^+$ without rearranging).

At the user side and since that the user is not all the time monitoring the network, we propose a new matrix P(*i*,*j*) with a numerical entry between 0 and 1 which represents the residual energy of each node. This matrix is updated if an unusual event happens like if there is a failed node and it is removed or a new node is inserted. The new matrix begins with P(*1*,*N*) where *N* is the number of nodes in the network then adding a new row at the top every time there is a new event. This matrix represents the history of the network and it is lasted as required (programmed) a day, a week or a month. In addition, this matrix can be 3D matrix where the third dimension represents the time at which that event happens.

In the previous example, the matrix at the beginning is $[1 \quad 0.8 \quad 1 \quad 0.1 \quad 0.4]$. Then after removing Node 4, the matrix will be $\begin{bmatrix} 1 & 0.77 & 0.9 & 0 & 0.395 \\ 1 & 0.8 & 1 & 0.1 & 0.4 \end{bmatrix}$. When the new node is inserted at the previous node's location, the matrix will be $\begin{bmatrix} 1 & 0.77 & 0.94 & 1 & 0.39 \\ 1 & 0.77 & 0.95 & 0 & 0.395 \\ 1 & 0.8 & 1 & 0.1 & 0.4 \end{bmatrix}$. The user is also able to see the topology matrix that was explained before.

V. Multi-Stack Layering Network

It is always difficult to isolate changes in any component of a system from its neighboring component within in its operation domain. Each component of any system is represented by stack and the interaction between each other is done in one or both directions as in Fig. 6. To maintain the laws of preservation of mass and energy, it is important to find the reaction of each component of the system.

The evaporation process is an example of a multi-stack layering network. In the hot days, water evaporates from any water body (i.e. lakes, ponds, seas,..) makes it decrease ($S1^-$). The water particles push the air particles to replace them in the atmosphere above the water body. Water is going up to condense in the upper cold layer forming clouds.

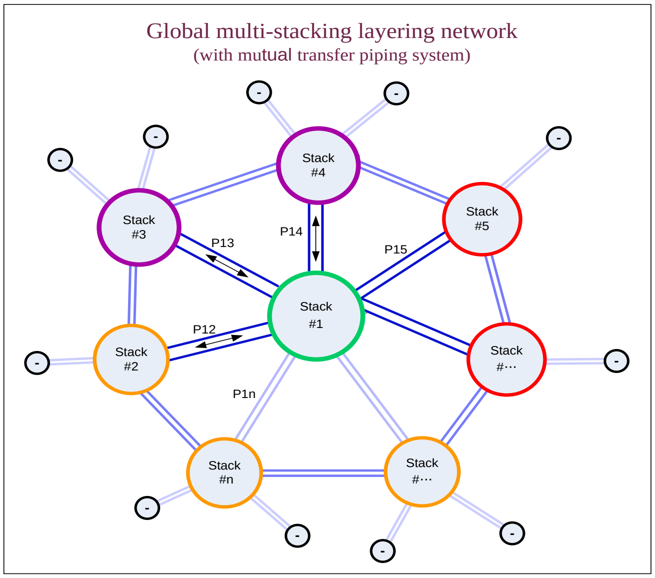
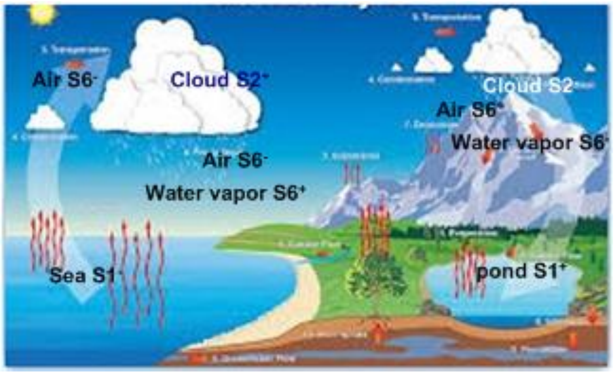
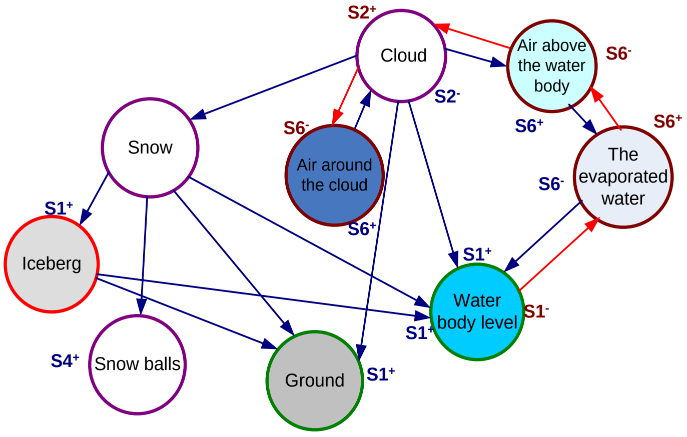


Fig.6. A general view of a multi-stack layering network [9].

The more water condenses the bigger or more clouds form. The cloud replaces the air that exists in this region pushing the air particles which they re-spread in this region. The opposite of the evaporation process is the rain in the worm regions. While if there is a cold region, the snow will be formed instead. In this case, iceberg could be formed or the snow will raise the sea level. If the snow falls on the ground, snowballs could be formed ($S4^+$) or the snow covers the ground ($S1^+$). This operation is a multi-stacking network as the system has many components and if anything happens to one of them, the rest reacts accordingly. This network is illustrated in Fig. 7. From this figure, we can say that all the multi-stack layering networks found anywhere are connected together to represent the universe.



(a)



(b)

Fig.7. General representation for the evaporation and condensation processes. (a) The evaporation and condensation processes. (b) multi-stack layering network in this processes.

In the case of WSN, as explained before there is a connection between the topology matrix, the routing table and the proposed history matrix as in Fig. 8. There are two options if the covering area is shrunk; either to insert the new node at the end of the network or insert it at the place of the failed one. In the first case (the left branch), the topology is changed as the nodes following the failed one are moved toward the sink then a new node is added at the end. The routing table is changed too with rearranging means all the following nodes will get one row up in the table after removing the row of the failed one (S6^{-}) adding a node at the end means adding a row at the end of the routing table (S2^{+}). In the second case, the failed node is removed from the network (S6^{-}) without changing the topology and in the routing table, the row of the failed one will be removed then adding a new node instead (S6^{+}) means adding a row in the routing table in the place of the removed one (S6^{+}) as in the right branch of Fig. 8.

VI. Conclusion and Future Works

It is crucial to search for new tools to analyze and model any system which provides us the chance for more understanding and deep control. The GDSP is one of them which helps us look at the system from different view and analyze it as layers connecting together. On the other hand, WSN is a rich application with many details which can be used to apply the new theory. A new matrix is proposed in this paper to retrieve and backward the history of the network so that the user follows the details of the network and does not miss any event. This history matrix is considered as a layer of the multi-stacking network in the WSN beside the routing table and the topology matrix.

What introduced in this paper is just opening the door for more works with more details about WSNs and trying to apply GDSP model from different respects. The required are; studying the effect of the new history matrix on the WSN and studying more systems and practical applications after applying GDSP to see the advantages of the new model where this paper is just putting a new theory with some applications without details.

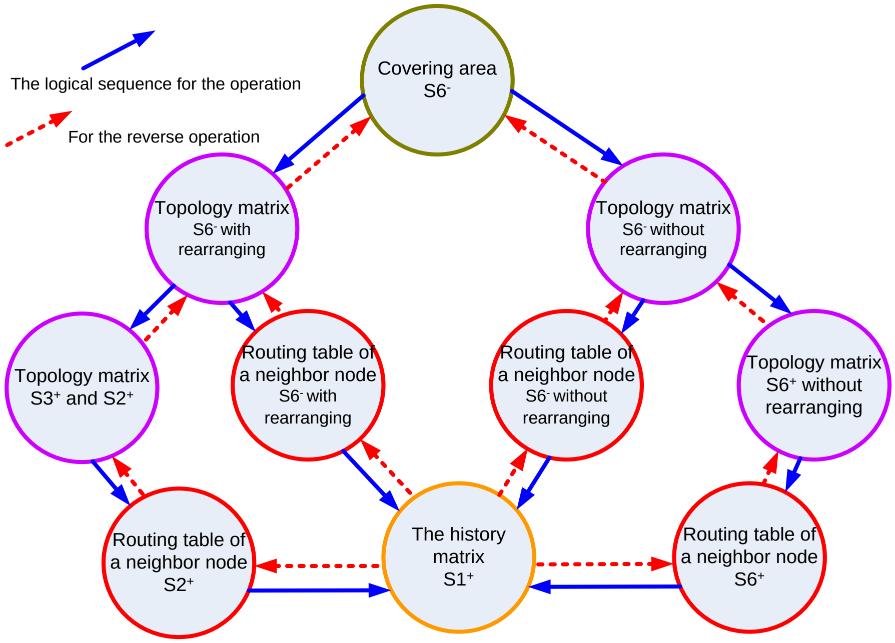

The logical sequence for the operation
For the reverse operation
Covering area $S6^{-}$
Topology matrix $S6^{-}$ with rearranging
Topology matrix $S6^{-}$ without rearranging
Topology matrix $S3^{+}$ and $S2^{+}$
Routing table of a neighbor node $S6^{-}$ with rearranging
Routing table of a neighbor node $S6^{-}$ without rearranging
Topology matrix $S6^{+}$ without rearranging
Routing table of a neighbor node $S2^{+}$
The history matrix $S1^{+}$
Routing table of a neighbor node $S6^{+}$

Fig.8. A multi-stacking network of WSN

REFERENCES

[1] *Y. Singh and A. Agrawal," Control systems, theory, problems and solutions," Galgotia publications Put. Ltd., India, 1st edn, 2008.*

[2] B. C. Kuo," Automatic control systems," Prentice-Hall of India, 7th edn., 1999.

[3] Ch. G. Cassandras and S. Lafortune," Introduction to Discrete Event Systems," Springer Science Business Media LLC, 2nd edn., 2008.

[4] R. Rizk, "Quality of Services in Wireless Sensor Networks. In Wireless Sensor Networks: From Theory to Applications, " Ibrahiem M.M.EI Emary and Ramakrishnan S. (Ed.), CRC Press, Taylor & Francis Group, USA, August 2013, http://www.crcpress.com/product/isbn/9781466518100.

[5] R. Rizk, S. Magdy, and F. Zaki, "Energy efficiency of virtual multi-input, multi-output based on sensor selection in wireless sensor networks," Wireless Communications and Mobile Computing, John Wiely & Sons, doi: 10.1002/wcm.2310, 2012.

[6] J. Zhang, X. Shen, G. Dai, Y. Feng, S. Tang, and C. Lv, "Energy-efficient lossy data aggregation in wireless sensor networks" , Adhoc & Sensor Wireless Net, vol 11, no. 1-2, pp. 11-35, 2011.

[7] J. A. Stankovic, "Research challenges for wireless sensor networks," Special issue on embedded sensor networks and wireless computing (SIGBED), vol 1, no.2, July 2004.

[8] Y. Pyokin, E. Jung, Y. Park, A radio-aware routing algorithm for reliable directed diffusion in lossy wireless sensor networks, Sensors, 2009, 9, (10), pp. 8047-8072.

[9] H. T. Dorrah, "Consolidity; Stack-based systems change pathway theory elaborated," Ain Shams Eng. J., 2014, http://dx.doi.org/10.1016/j.asej.2013.12.002

[10] H. T. Dorrah, "Supplement to consolidity: moving opposite to built-as-usual practices, " Elsevier Ain Shams Eng. Journal, vol 4, no. 4, pp. 783-803, 2013.

[11] H. Elhadidy, R. Rizk, and H. T. Dorrah, "A new generalized data stacking programming (GDSP) model", Proc. of 9th ICENCO Egypt, Dec. 2013, pp. 78-84, DOI:10.1109/ICENCO.2013.6736480.

[12] J. Mireles and F. Lewis, "Intelligent material handling: development and implementation of a matrix-based discrete event controller," IEEE Transactions on Industrial Electronics, vol. 48, no. 6, pp. 1087– 1097, Dec. 2001.

[13] D. Goldenberg, J. Lin, A. Morse, B. Rosen, and Y. Yang, "Towards mobility as a network control primitive," proc. of the ACM Int. Symp. on Mobile Ad Hoc Net. and Comp. (MobiHoc), pp. 163-174, May 2004.

[14] L. Lazos, R. Proovendran, and J. A. Ritcey, "Probabilistic detection of mobile targets in heterogeneous sensor networks," proc. of the 6th int. Conf. on Information Processing in Sensor Net. (IPSN'07), pp. 519-528, April 2007.

[15] R. Rizk, H. Elhadidy and H. Nassar, "Optimised mobile radio aware routing algorithm for wireless sensor networks," IET Wirel. Sens. Syst., vol 1, no. 4, pp. 206–217, 2011.

[16] M. Chen, T. Kwon, and Y. Choi, "Energy-efficient differentiated directed diffusion (EDDD) in wireless sensor networks," Elsevier Computer Comm., special issue on Dependable WSNs, vol. 29, no. 2, pp. 231-245, 2006.

Data Protection Control and Learning Conducted Via Electronic Media I.E. Internet

Mohamed F. AlAjmi,
Associate Prof., King Saud University
Riyadh, Saudi Arabia

Shakir Khan (Corresponding Author)
Researcher, King Saud University
Riyadh, Saudi Arabia Arabia
Nationality
Indian

Abdulkadir Alaydarous
Associate Prof., Faculty of science
Taif University, Taif, KSA

***Abstract*—numerous e-taking in establishments are hurrying into receiving ICT without precisely arranging and seeing any related security concerns. E-learning is another technique for taking in which eventually relies on upon the Internet in its execution. The Internet has turned into the venue for another set of unlawful exercises, and the e-taking in environment is notwithstanding laid open to such dangers. In this paper, e-learning setting, definition, aspects, improvement, development, profits and tests are all explained upon. This paper examines the security components needed to be executed inside e-learning situations. Moreover, the paper clarifies the circumstances and existing examination relating to security in e-taking in. Besides, data security administration is inferred to help setting up a secured e-taking nature's turf.**

Keywords—e-Learning; e-taking; ICT; data security component

I. Introduction

The development of Information and Communication Engineering (ICT) has critical impacts on all individuals around the globe. With this development, individuals have the capacity to associate with one another, particularly through the Internet. Nowadays, the Internet itself is definitely shifting the procurements of administrations and merchandise, basically in view of its characteristics: promptness, openness, pervasiveness, and worldwide scope. Besides, eservices have been presented broadly; consequently, the instruction industry has completely kept its new potential as long life taking in instruments from the Internet characteristics, for example, as the web requisition, case in point. This industry is ready to turn into one of the biggest segments on the planet economy. The advancement of e-taking in has consequently prompted another method for taking in and, in the meantime, has given equivalent chances to everybody to end up learners. With such routines for taking in now accessible, it is said that data or information might be arrived at fingertip level, and consequently empower scholars to outperform in their studies. In any case, notwithstanding the Internet as a spot to acquire all fundamental data and learning, it has additionally turned into the venue for another set of illicit exercises. Data on the Internet is persistently presented to security dangers. As an outcome of e-taking in needing to rely on upon the Internet or, particularly, for the most part through web provisions, the e-taking in environment has likewise get influenced by security dangers. With this in attention, this paper intends to investigate the more extensive setting of data security issues and dangers, furthermore the potential of data security administration in lessening them. The main some piece of this paper talks about the e-learning setting — the definition, trademark, improvement, development, profit and the tests — all of which think about security in e-taking in as another challenge in executing the e-taking in nature's domain. Additionally, web requisitions are the medium used to support the larger part of online administrations furthermore, henceforth, have turned into the prime focus of Internet strike.

The second a piece of this paper takes a gander at data security in e-taking in which has been dismissed in research. Numerous e-taking in establishments are surging into receiving ICT without deliberately arranging and comprehending the ever-display security concerns. Issues, for example, real clients, course content unwavering quality, and receptiveness (counting the suitability and accessibility), and in addition other contemplations, all necessity to be deliberately tended to in request to guarantee the taking in procedure can successfully occur. In conclusion, the paper will examine the potential of data security administration to be executed in the connection of e-taking in, with a specific end goal to set up a secured e-taking nature's domain.

II. Internet Learning

E-taking in is the term used to depict the utilization of the web and other Internet advances regarding upgrading the showing and taking in experience. It offers comparative attributes of numerous different eservices, such e-business, e-saving money and e-government. The e-administrations clients' behaviors are diverse as per their parts and necessities. E-learning clients concentrate on how to profit from e-learning concerning showing and taking in purposes. The clients may need to invest longer times of time the point when undertaking e-taking in contrasted with different eservices.

A. Classification And Distinctiveness

There are numerous distinctive meanings of e-taking in exhibit in the expositive expression, and everyone has an alter accentuation: a few concentrate on the substance, some on correspondence, and some on the innovation [1]. One of the unanticipated definitions for e-taking in was gave by the American Society to Training and Improvement (ASTD), which suggests that e-learning blankets a wide set of provisions and procedures, for example, electronic taking in, computer based taking in, virtual classrooms, and digitals joint effort.

E-taking in is the usage of innovation in request to help the taking in procedure, whereby learning or data might be entered utilizing the correspondence innovation. The taking in methodology can be consistent, given that the substance is accessible on the net. [2] characterizes e-taking in as a part of adaptable taking in, which is a wide situated of requisitions and methods, all of which utilize all accessible electronic media to convey instruction and preparing; this incorporates machine based taking in, web-based taking in, virtual classrooms, and advanced coordinated effort.

B. E-taking in Development and Growth

The utilization of innovation to help taking in was began as unanticipated as the 1980s. Such an improvement was likewise in conjunction with the dispersal of PCs for particular utilization around then. Actually, higher taking in establishments have likewise drastically changed in the course of the most recent thirty years in light of strategy drivers, for example, broadening interest, long life taking in, and quality certification [3], Figure 1 underneath shows the development of e-taking in through 1983 until the present day. The accentuation on e-taking in the past has been on the 'e', which alludes to electronic or innovation. There is an urge to movement to the taking in (substance) in guaranteeing the achievement of e-learning. In addition, there are some normal terms which are utilized conversely to reflect the use of innovation in instruction, for example, dispersed instruction, e-taking in, separation training, mixed taking in and online classes. Separation instruction relates more to taking toward oneself in. In this occasion, the takings in material are posted through physical mail or could be entered on the web. The meeting sessions are directed just a couple times for every semester. Then, the synthesis of face-to-face furthermore web taking in sessions is alluded to as mixed taking in and are truly prominent these days.

It is a technique for instructing at a separation which utilizes engineering joined with customary training or preparing. Vital taking in conveyance channels are utilized, for example, physical classrooms, virtual classrooms, print, email and message sheets, tutoring frameworks, programming reproductions, online joint effort, and versatile and remote channels [5].

As delineated in Figure 2, [1] position e-taking in as a kind of separation training. They likewise specify that 'conveyed instruction' is a more extensive term which incorporates parts of separation and online training and additionally being mixed with face-to-face taking in.

Pre1983 - Era of Educator headed Preparing	This was the predominant showing instrument before Pcs got to be generally accessible, and when connections between the teacher and people occurred in the classrooms.
1984-1993 – Multimedia	Windows 3.1, Macintosh and CD Roms were the fundamental innovation improvements throughout this period. On the other hand, classroom connections and element presentations needed in this medium.
1994-2000 – network Immaturity	As the web advanced, the entry of email, media players and streaming audio/video started to change the substance of media mediums. Learners were equipped to gain access to address notes or materials from the web at whenever and at any (Internet-able) area.
2001 and beyond – Next-generation Web	Progressed site plan, rich streaming media (genuine audio/video) and high data transmission (speedier information stream) will revolutionize the path in which instruction will be conveyed. Teacher headed, intelligent modes can now happen through the web, arriving at significantly a larger number of scholars than in the recent past.

Fig. 1. The Maturity of E-taking in [4]

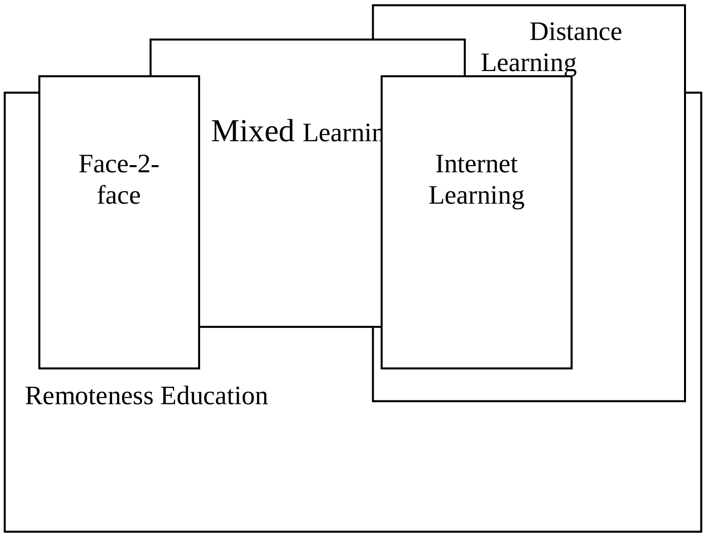


Fig. 2. The Relationship of E-figuring out how to Distribution Learning [1]

Today, the usage of e-taking in is a combo of three methods for utilizing innovation: utilizing engineering non-concurrently, i.e. just as instruments to help or supplement a customary taking in, utilizing engineering non-concurrently and synchronously as devices to help or supplement a universal (eye to eye) taking in, and utilizing engineering non-concurrently and synchronously to convey a taking in course (totally online). Figure 3 delineates the stream of correspondence between people and teachers in the e-taking in framework segment.

Despite the claim that numerous e-taking in activities have missed the point of desires, the business sector of e-taking in is all things considered proceeding to develop. Reports from the Sloan Foundation show that 3.5 million people (speaking to just about 20% of all U.S. higher training people) enlisted in at minimum one online course throughout the fall 2007 term [6]

This development is fuelled by new establishments entering into the online stadium, joined with a consistent person interest for web taking in alternatives. The necessity of learning specialists has additionally helped the development of e-taking in: every representative necessities to furnish themselves with the learning and aptitudes to as incredible a degree as could reasonably be expected, with the goal that they can advance; the simplest approach to do this is to enroll as an e-taking in learner.

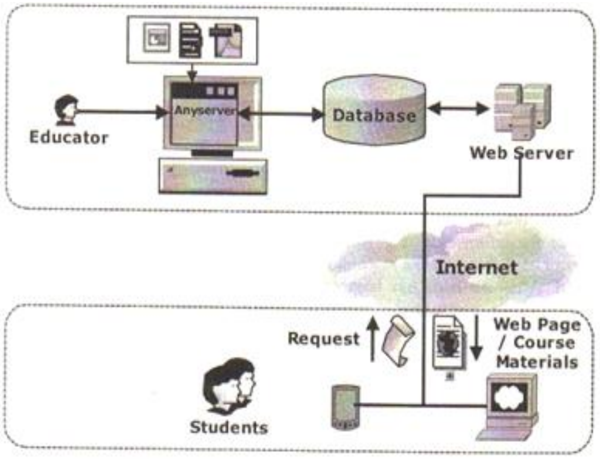


Fig. 3. E-Learning Systems Diagram

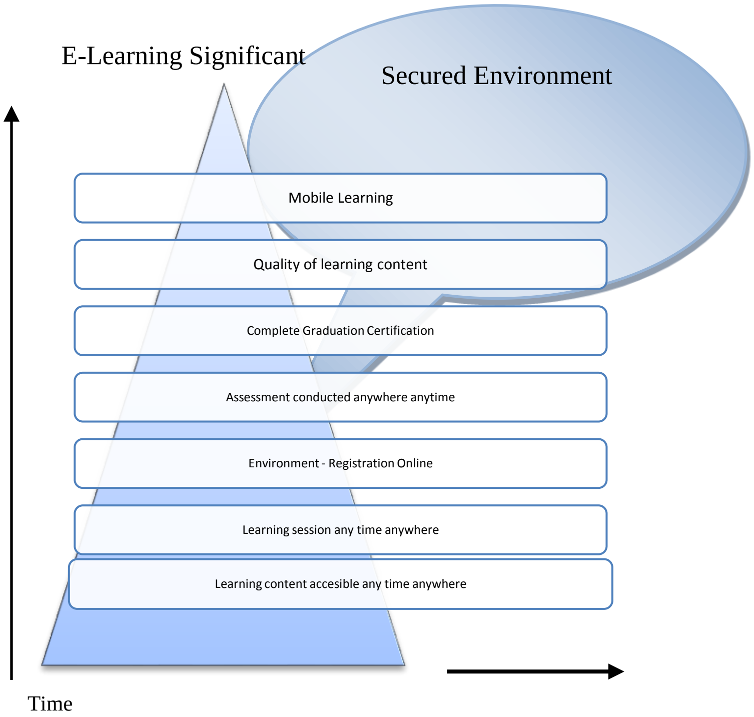

E-Learning Significant
Secured Environment
Mobile Learning
Quality of learning content
Complete Graduation Certification
Assessment conducted anywhere anytime
Environment - Registration Online
Learning session any time anywhere
Learning content accesible any time anywhere
Time

Fig. 4. The E-Learning Functionality Growth

The practically of e-taking in has additionally developed in parallel with the requirements and the improvements of innovation. Figure 4 shows the development in e-taking in practicality. At first, e-taking in distributes the taking in substance on the Internet, and empowers it to be approachable by the client at whatever time and at any area (non-concurrent taking in). It then increases to permit the taking in session to be led at whatever time and anyplace (synchronous taking in) from the points of view of all clients (instructor/lecturer and learners). E-taking in now empowers the enrollment, appraisal, and posting graduation accreditation on the web. With the plan of including more terrific adaptability, portable taking in has been presented in spite of the way that utilization is as of now constrained and not completely utilized. As the practicality of e-taking in keeps on growing, the e-taking in environment needs to end up additional secured. More terrific usefulness introduced to clients will make the e-taking in environment more open and presented to the data security dangers.

III. Advantages of E-Learning

These days, workers must have the capacity to accompany and stay in-accordance with mechanical progressions and in like manner perform creative critical thinking. One method for taking care of the demand for these new abilities, particularly in Information Technology, is by means of e-learning, which likewise offers the potential for persistent taking in.

E-taking offers everybody the chance to turn into a learner. The idea of at whatever time, anyplace taking in advertises long lasting taking in and in like manner wipes out the issues connected with separation. The adaptabilities which e-taking offer to the understudies are the primary spurring component in picking online courses [7]. Also, the use of engineering in taking in will give different other points of interest, for example, enhancing the nature of taking in, enhancing access to training and preparing, diminishing the expenses connected with training, and enhancing the expense viability of instruction. E-learning gives a stage of an overall outlined, learner-centered, captivating, intelligent, reasonable, productive, effectively approachable, adaptable, and seriously dispersed and encouraged e-taking nature. Also, learners can spare cash what's more of a chance used on voyaging and getting the right materials for their study. They can diminish printing sets back the all finances by perusing the accessible taking in materials on the web. Moreover, e-taking in increments access to taking in materials. It likewise empowers scholars to have more extensive access to constrained assets, for example, e-diaries furthermore e-books. This can help the scholars in upgrading their taking in. By taking out obstructions of time, separation and socio-investment status, people can now assume responsibility of their own long lasting taking in. The enhanced correspondence connection and better understudy access energizes enhanced cooperation. Understudies can have open discussions, permitting them to correspond with their companions, or even private gatherings between the learner and the instructor or educator. An alternate profit offered by e-taking in is quicker conveyance of appraisals, as instructors can give reaction speedier contrasted and the universal system and learners can likewise help criticism around themselves.

A. *Confronts in E-Learning*

Executing e-taking in is not a simple assignment. Regardless of numerous profits picked up from e-taking in, there are likewise issues and tests when intending to make e-taking in fruitful. The tests, as reflected in Figure 5, are recognized from two points of view: the taking in supplier, and the client. From the taking in supplier view, Higher Learning Institutions (HLI) are encountering troubles in connection to different innovative issues, for example, planning productive framework. Transmission capacity and connectivity are eventually fundamental, since people will be reliant on these offices to gain access to taking in materials on the web. Besides, the conveyance of high transmission capacity substance, for example, advanced film, is still hazardous to the home client. Taking in material is likewise an issue, since an absence of value substance is planned. Creating great substance for people might as well think about numerous diverse variables, for example, pedagogical viewpoints, human-PC interface, and smoothness. Guaranteeing these are generally readied obliges a high plan; therefore, high expenses for usage are not out of the ordinary. In a creating nation, these tests are significantly more troublesome, essentially due to the assets issues confronted.

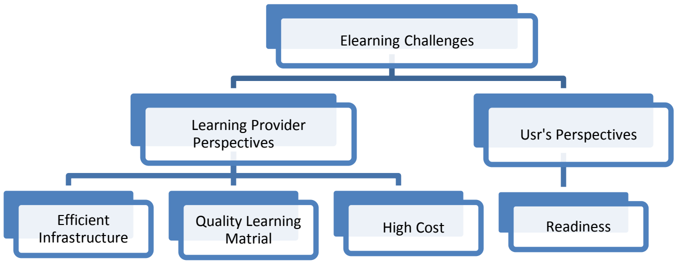


Fig. 5. The E-taking in Challenges

From the clients' viewpoints, they are encountering challenges in the settings of status. (Aziz) proposes basic components in planning individual's status incorporate duty and aptitudes. Preparation incorporates status of information, in addition to availability of inspiration for taking toward oneself in. Scholars are most certainly not readied for e-taking in due to low workstation proficiency and low levels of order toward oneself for self learning techniques. Moreover, consistent with the Engineering Acceptance Model (TAM), the recognized advantage and discerned usability do have affects on clients' acknowledgement on the engineering utilization: assuming that they don't perceive how e-taking in can offer assistance them, the person will at last oppose continuation, then again even enroll on the grounds that they think they might come up short due to absence of help and preparing gave by the taking in supplier. Besides, teachers might likewise feel the same; accordingly, an alternate purpose behind not having any desire to utilize e-taking in is in light of the fact that they see little remunerate. The development of e-learning is fuelled by new foundations entering the online coliseum joined with a proceeded person interest for web taking in choice. Around the components to be recognized when creating the e-learning nature's domain, mixed media direction, independent taking in, educator headed association, and change of taking in viability, and social vicinity are all contemplations. These tests are by one means or another related when making secure and fruitful e-taking in situations in the feeling of classifiedness, accessibility and trustworthiness. Essentially, data security in e-taking in is a challenge which is infrequently examined. Security in e-learning has been slighted and deserted [9].

B. *INFORMATION INTIMIDATION ON WEB*

With attention to ICT, individuals these days are getting the profits of entering unfathomable data rapidly. Data may exist in numerous structures: it can be printed or composed on paper, archived electronically what's more transmitted by post or by electronic methods. Whatever structure data takes or the methods by which it is imparted, it may as well dependably be properly secured.

Data inferring from handy information is around an organization’s principle holdings. All things considered, when it is dependably simple for everyone to gain access to, it will accordingly additionally be simple and advantageous for anyone to addition access, regardless of if they have great or terrible proposition. As an aftereffect of this expanding interconnectivity, data is currently presented to a developing number and a more extensive mixed bag of dangers and vulnerabilities. Accordingly, data must be ensured keeping in mind the end goal to dodge the misfortune of its privacy, respectability and accessibility. Some of the most genuine dangers are recorded as beneath.

- Intentional programming strike (infections, worms, macros, refusal of administration)
- Specialized programming disappointments and slips (bugs, coding issues, obscure tricks)
- Demonstrations of human slip or disappointment (mischance’s, representative missteps)
- Conscious demonstrations of reconnaissance or trespass (unauthorized access or information accumulation)
- Planned demonstrations of treachery or vandalism devastation of data or framework)
- Specialized fittings disappointments or failures (supplies disappointment)
- Planned demonstrations of burglary (unlawful appropriation of supplies or data)
- Bargains to protected innovation (theft, copyright, encroachment)
- Nature of Service deviations from administration suppliers (power and WAN administration issues)
- Mechanical oldness (obsolete or out-dated innovations)
- Purposeful demonstrations of data blackmail (extortion for data exposure).

C. Data Security in E-Learning

E-taking in is primarily subject to data as well as correspondence innovations. As per [10] e-taking in is dependent upon three basic criteria, which are: 1) system proficient redesigning, storage/retrieval, circulation and imparting of data; 2) conveyance to the finish of client through machine utilizing standard Internet innovation; and 3) keep tabs on the broadest perspective of e-taking in. The principal also second criteria uncover the e-taking in foundations to the dangers, as the utilization of ICT could at last lead to numerous conceivable data security dangers which could trade off data, for example, misfortune of classifiedness, accessibility, introduction of basic information, what's more vandalism of open data administrations [11] Shockingly, not many endeavors have been made to amend this circumstance. More exertions have been emphasized to improve the substance and innovation because of tending to substance and innovation as the challenges in securing a great e-taking in nature.

Security is required inside e-taking in situations owing to the way that, these days, information has turned into a significant method of generation, as item and as a key for individual victory. In e-taking in, data inferring from advantageous information is around the primary stakes of the organization. Around security issues in e-taking in are insurance against control (learners, insider), client confirmation, and secrecy [11]. In any case, as the practicality of e-taking in is stretching, data must be eagerly ensured in this greater connection to keep away from the misfortune of its classifiedness, uprightness and accessibility. Some individuals may state that information ought to be imparted, be that as it may there are circumstances where the stream of touchy data ought to be limited to just a couple of well-defined assemblies, for example, for instance, taking in materials for specific assemblies and copyright security of educated properties. Moreover, it is troublesome to confirm whether a task has been finished and sent by a substantial learner. The character what's more the protected substance are troublesome to look after. E-taking portions comparable aspects of other e-administration. There are three primary qualities of each e-benefit: the administration is approachable through the Web, the administration is devoured by an individual through the Web, and there could be a charge which the purchaser pays the supplier for utilizing the e-administrations. The practicality and security dangers to e-taking in have normal characteristics with other e-administrations, and the administration methodologies could likewise have normal qualities. Assuming that organizations are to secure and maximize the profit for their speculation in taking in engineering, substance and administrations, the frameworks they use must be interoperable, usable, sensible, and strong [12] Past studies have indicated that there are boundaries to an all the more across the board selection of online training [6]. The purpose for such restraints is most certainly not the high expenses or the more excellent level of undertakings which necessity to be done, but instead the security viewpoint, which is something that is truly immaterial in the digital world. Eventually, it is troublesome to check whether the task is finished and sent by a honest to goodness person. The character and the safe substance are troublesome to support. Besides, security issues in e-taking in have been tended to basically by security engineering; for instance a specialized schema on validation and responsibility, access control, ensure of interchanges, non-revocation issues and taking in asset supplier server assurance (Furnell 2001).

D. Data Security element in E-learning

Data security is the assurance of data from dangers. It is executed in place to guarantee business progression and to as needs be minimize business hazard. In the meantime, it is wanted that there be a great profit for ventures furthermore business chances. The e-taking in points are concerned with giving showing and e-figuring out how to everybody.

Guaranteeing the accessibility and respectability of data is the fundamental objective in connection to e-taking in security. Accessibility in e-taking in is the confirmation that the e-taking in environment is open by authorized clients, at whatever point required. Two aspects of accessibility are regularly talked about, which are disavowal of administration and misfortune of information transforming proficiencies. The e-taking in clients are reliant on the data on the Internet; hence, the accessibility of materials and data to be entered at any time and any area is urgent. Neglecting to fulfill this will have a tremendous effect on e-taking in clients and e-learning suppliers. [14] specify that a few characteristics which influence e-taking in are security furthermore security for e-conveyance and shared training. The accessibility of materials and data is deficient. It is imperative to assurance the unwavering quality of the materials and the data distributed. This identifies with an alternate security component, which is respectability. Trustworthiness in e-taking in is the security of information from purposeful or inadvertent unauthorized changes. Respectability relies on upon access controls; hence, it is important to decidedly and interestingly distinguish all persons who endeavor access. Trustworthiness can be bargained by programmers, impostors, unauthorized client movement, unprotected downloaded records, Lans, and unauthorized projects (e.g., Trojan stallions and infections), essentially since each of these dangers can prompt unauthorized progressions to information or programs. Despite the fact that accessibility and respectability are the fundamental security components which oblige attention inside e-taking in situations, the component of privacy is likewise imperative. Privacy is the assurance of data in the framework so that unauthorized persons can't get access. The emulating are probably the most generally experienced dangers to data privacy: programmers, impostors, unauthorized client movement, unprotected downloaded records, neighborhood territory systems (Lans), and Trojan stallions.

E. Examine in safety measures of E-Learning

Securing the e-taking in environment requires staying away from the four sorts of risk, which are manufacture, adjustment, interference and capture. Right now, little research has been directed to secure the e-taking nature's turf.

Looks into in security mostly concentrate on three fundamental regions: arrangement, character (which alludes to gain access to administration) and licensed innovation. Most scientists state that, with a specific end goal to stay away from all assaults upon the e-taking nature's domain, regulating access is foremost. One of the approaches to do this is by means of verification and authorization process. [15] suggests a confirmation transform to recognize a lawful client transform; this will defeat the illicit use of provision. A framework which is as well intensely secured will be challenging to be entered by the client. To adjust access and security, Saxena (2004) notice giving clients single sign on verification and authorization administrations to all authorized web requisitions and web assets. [11] recommends a methodology to ensuring educated property by enlarging the control of the copyright holder on the whole lifetime of the advanced information. He recommends a technique reputed to be CIPRESS, which controls the right to gain entrance to the material. Yong [16] examines an alternate specialized perspective concerning how to secure e-taking in by computerized personality outline and security protection. Nonetheless, regulating gain access to by utilizing certain engineering mechanisms is recognized lacking, subsequent to the ambush does not so much hail from pariahs yet could additionally hail from the insider. The correct supervision of the treatment of data security issues is imperative to guarantee no vulnerabilities. Along these lines, the data security administration is imperative when striving to guarantee the accomplishment of secured e-taking in execution.

IV. Conclusion

E-taking in has developed and is extending at an extremely quick pace. The profits it offers build the number of e-taking in clients. The practicality of e-taking in keeps on growing and depend an increasing amount intensely on the Internet. On the other hand, the Internet has turn into a position of illicit exercises, which subsequently uncover e-figuring out how to dangers. Guaranteeing the accessibility also respectability of data and material inside e-learning situations obliges that counter measures, for example, security engineering equipment and programming, requirement to be executed. In any case, it is acknowledged deficient. Additionally, IMS is required in request to guarantee the security of the e-taking in nature. ISM for e-taking in is no diverse to other e-administrations; in any case, due to the adaptability component offered by e-taking in and diverse client behaviors, e-taking in obliges a security administration skeleton which can go about as an aide in helping the e-taking in supplier (organizations) in dealing with the data security inside the e-learning nature. Moreover, the combo of ISM and the present data security innovation utilized will give better comes about within the achievement of security execution.

Acknowledgment

The authors acknowledge the support from Research Center, College of Pharmacy, King Saud University, and Riyadh (KSA).

References

[1] Mason, R. and Rennie, F. (2006), E-learning: the key concepts, Routledge, Abingdon Great Britain.

[2] Eklund, J., Kay, M. and Lynch, H. M. (2003), E-learning: emerging issues and key trends: A discussion paper, Australian National Training Authority, Australia.

[3] Conole, G., Smith, J. and White, S. (2007), 'A critique of the impact of policy and funding', in Conole, G.and Oliver, M. (eds.) Contemporary perspectives in E-learning Reserach themes, methods and impact on practice, Routledge, London; New York, pp.38-54.

[4] Dietinger, T. (2003), Aspects of E-Learning Environments (unpublished Doctor of Technical Sciences thesis), Institute for Information Processing and Computer Supported New Media (IICM), Graz University of Technology, Austria.

[5] Morrison, D. (2003), E-learning strategies, Wiley Chichester.

[6] Allen, E. and Seaman, J. (2007), Online Nation Five Years of Growth in Online Learning, 1, Sloan Consortium, United States.

[7] Jain, K. K. and Ngoh, L. B. (2003), 'Motivating Factors in e-learning -a Case study of UNITAR',Student Affairs Online, [Online], vol. 4, no. 1, pp.21, June, 2008 available at: http://www.studentaffairs.com/ejournal/Winter_200 3/e-learning.html.

[8] A. Aziz, S. H., M.Yunus, A. S., A. Bakar, K. and B.Meseran, H. (2006.), 'Design and development of learning management system at university Putra Malaysia: a case study of e-SPRINT. I', WWW 06: Proceedings of the 15th international Conference on World Wide Web, May 23 - 26, 2006, Edinburgh, Scotland, ACM, New York, pp.979-980.

[9] Raitman, R., Ngo, L. and Augar, N. (2005), 'Security in the Online E-Learning Environment', Advanced Learning Technologies, 2005.ICALT 2005.Fifth IEEE International Conference on Advanced Learning Technologies, pp.702-706.

[10] Rosenberg, M. J. (2001), E-learning strategies for delivering knowledge in the digital age, McGraw-Hill, New York.

[11] Graf, F. (2002), 'Providing security for eLearning', Computers & Graphics, vol. 26, no. 2, pp.355-365.

[12] Norman, S. and Da Costa, M. (2003),'Overview of e-learning Specifications and Standards', Open Learning Agency, and Eduspecs Technical Liaison Office.

[13] Furnell, S. M. and Karweni, T. (2001), 'Security issues in Online Distance Learning', VINE: The Journal of Information and Knowledge Management Systems, vol. 31, no. 2.

[14] Yang, C., Lin, F. O. and Lin, H. (2002), 'Policy based Privacy and Security Management for Collaborative E-education Systems', Proceedings of the 5th IASTED International Multi-Conference Computers and Advanced Technology in Education (CATE 2002), pp.501–505.

[15] Dr. AlAjmi, Dr Shakir Khan, Using Instructive Data Mining Methods to Revise the Impact of Virtual Classroom in E-Learning"International Journal of Advanced Science and Technology"Vol. 45, August, 2012"

[16] Yong, J. (2007), 'Digital Identity Design and Privacy Preservation for e-Learning', Proceeding of the 2007 11th International Conference on Computer Supported Cooperative Work in Design , pp. 858-863.

A Code Level Based Programmer Assessment and Selection Criterion Using Metric Tools

Ezekiel U. Okike
Department of Computer Science
University of Botswana
Gaborone, Botswana

***Abstract*—this study presents a code level measurement of computer programs developed by computer programmers using a Chidamber and Kemerer Java metric (CKJM) tool and the Myers Briggs Type Indicator (MBTI) tool. The identification of potential computer programmers using personality trait factors does not seem to be the best approach without a code level measurement of the quality of programs. Hence the need to evolve a metric tool which measures both personality traits of programmers and code level quality of programs developed by programmers. This is the focus of this study. In this experiment, a set of Java based programming tasks were given to 33 student programmers who could confidently use the Java programming language. The codes developed by these students were analyzed for quality using a CKJM tool. Cohesion, coupling and number of public methods (NPM) metrics were used in the study. The choice of these three metrics from the CKJM suite was because they are useful in measuring well designed codes. By examining the cohesion values of classes, high cohesion ranges [0,1] and low coupling imply well designed code. Also number of methods (NPM) in a well-designed class is always less than 5 when cohesion range is [0,1]. Results from this study show that 19 of the 33 programmers developed good and cohesive programs while 14 did not. Further analysis revealed the personality traits of programmers and the number of good programs written by them. Programmers with Introverted Sensing Thinking Judging (ISTJ) traits produced the highest number of good programs, followed by Introverted iNtuitive Thinking Perceiving (INTP), Introverted iNtuitive Feelingng Perceiving (INTP), and Extroverted Sensing Thinking Judging (ESTJ)**

Keywords—computer programs; program quality; class cohesion; programmers; personality traits

I. Introduction

Programming is a challenging task, which requires appropriate skills as well as appropriate temperamental suitability. Among the skills often demonstrated by professional and successful programmers are logical and analytical thinking, problem understanding and interpretation, detailed understanding of a programming language's syntax and a good communication ability. Capretz and Ahmed [1] identified some of the skills required in computer programming to include strong analytical and problem solving skills, communication skills, interpersonal skills, ability to work independently, active listening skills, innovative skills, organizational skills, openness and adaptability skills, fast learning skills and team playing skills. Apart from possessing these skills, success as a computer programmer may also be influenced by personality types such as Extroversion (E), Introversion (I), Sensing (S), iNtuition(N), Thinking(T), Feeling(F), Judging(J) and Perceiving(P) (Okike and Olanrewaju[2]; Capretz and Ahmed [1]; Capretz[5]; Da Cunha and Greathead[3]; Tueley and Bieman [6]; Bentley [4]).

Furthermore, the reliability of acomputer software depends on the code level quality of the program which indeed results from programmers coding skill. For this reason, it becomes very necessary to evolve code level measurement of program quality and individual programmer personality traits in the selection process of career computer programmers. Hence matching coding skill with personality traits will enable the identification and selection of good computer programmers. This is the motivation for this paper.

A. Problem Statement

Programmers are widely perceived as Introverts, Sensors and Thinkers (Capretz and Ahmed [1], Sensing and iNtuitionist, (Da Cunha and Greathead [3], Introverts, iNtuitionists, Thinkers and Judges (INTJ) (Tieger, [20]). These assessments are purely based on personality traits factors without recourse to code level quality of programs or resulting software. The present study seeks to bridge the gap between programmer personality traits and the quality of programs written by programmers by making use of a two level metrics based on both personality traits and code level quality to assess and to select competent programmers who create quality software programs. The Quality of a Program (QoP) in this study is measured in terms of the Cohesiveness of the Program module (CoPm), Coupling Between Object classes (CBO) and Number of Public Methods (NPM). In software development, high cohesion (range [0,1]) and low coupling imply good design. In addition, number of methods (NPM) in a well-designed class is always less than 5 when cohesion is high, range[0,1] and coupling is low [17]. The cohesion degree of a component is high if it implements a single logical function, and cohesive component tend to have high maintainability and reusability ((Okike [7], Badri [8], Bieman and Kang[9])

B. Study Ovjectives

The main objective of this study is to create a two level metrics which is based on programmers personality traits and the code level quality of program modules. This instrument should be useful in selecting programmers who create quality programs. Specifically, the objectives of this study are to:

- investigate the personality traits of skilled programmers using Myers Briggs type indicator (MBTI)
- investigate the code level quality of programs written by programmers using Chidamber and Kemerer Java Metric tool (CKJM)
- suggest the personality type indicator(s) of competent programmers.

C. *Research Questions*

The following research questions are investigated in this study.

- What are the personality traits of good computer programmers?
- Which personality traits designed quality (cohesive) programs?

D. *Research Hypotheses*

The following hypotheses are tested in this study:

- H1: Introverts design better codes than extroverts in terms of class cohesion

H0: Introverts do not design better codes than extroverts

- H1: Sensors design better codes than intuitives in terms of class cohesion

H0: Sensors do not design better codes than intuitives

- H1: Thinkers design better codes than feelers in terms of class cohesion

H0: Thinkers do not design better codes than feelers

- H1: Judges design better codes than Perceivers in terms of class cohesion

H0: Judges do not design better codes than perceivers

- H1: There is significant correlation between personality traits and code quality

H0: There is no correlation between personality traits and code quality

The rest of this paper is divided into 7 sections. Section 2 is a presentation of the conceptual model of the study. Section 3 is the literature review. Section 4 explains the research methodology. Section 5 presents the result of this study with appropriate discussion. Section 6 is the conclusion while section 7 is the list of references

II. Conceptual Framework

The framework for this study is based on Capretz and Ahmed [1] model : Mapping programmers and skills to personality type as shown in figure 1 below and Okike [7] Metric calculation Process using Chidamber and Kemerer metric tool as shown in figure 2. Arising from these two models is a hybrid adapted from the two to achieve the objectives stated in section 1.3.

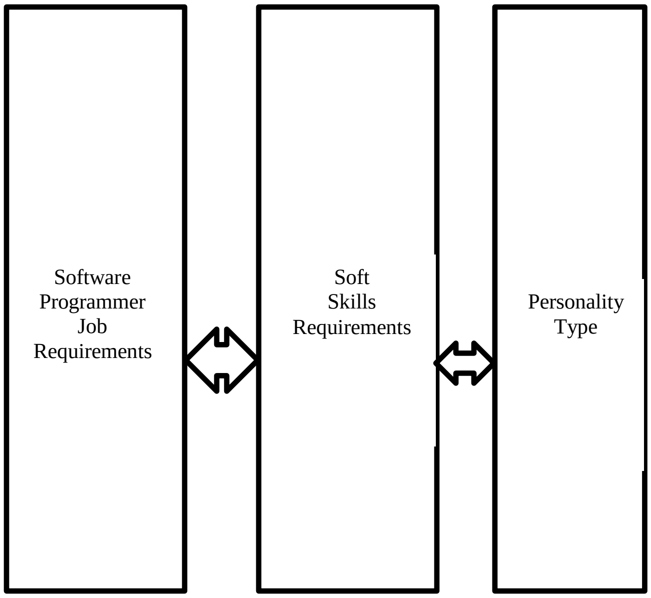


Fig. 1. Mapping Programmers and skills to personality type Adapted from Capretz and Ahmed [1]

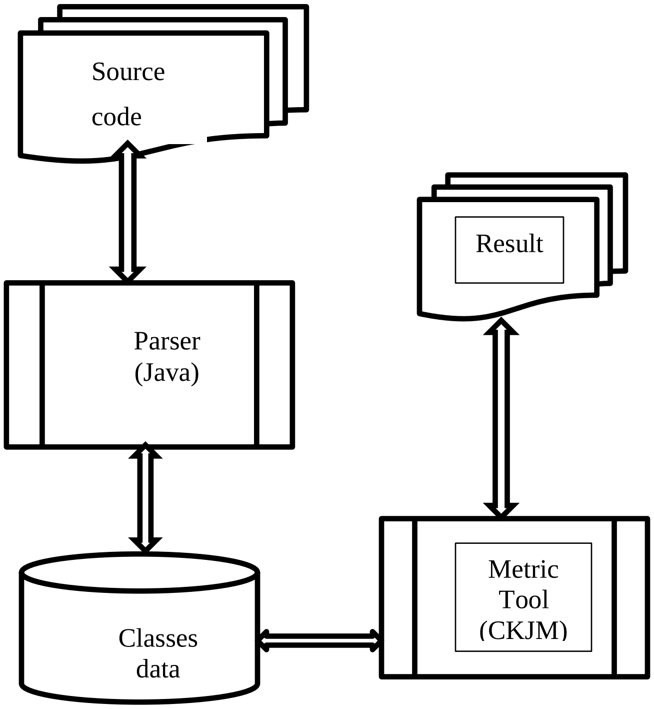


Fig. 2. Metric calculation processSource: Adapted from Okike [7], and Badri [8]

A. *Proposal of a two level Metrics Model for selecting programmers*

Figure 3 below shows a hybrid model metric tool for selecting programmers who create cohesive software. Since code level quality is measured by high cohesion, any metric tool which measures cohesion in software would be good candidate. In this study, the Chidamber and Kemerer Java Metric (CKJM) tool is used with particular focus on the Lack of Cohesion in Methods (LCOM) metric. The hybrid tool has 2 levels as shown in figure 3 namely: level 1 – programmer personality trait measurement using the Myers Briggs Type Indicator (MBTI) and level 2- program quality measurement using Chidamber and Kemerer Java metric (CKJM) tool.

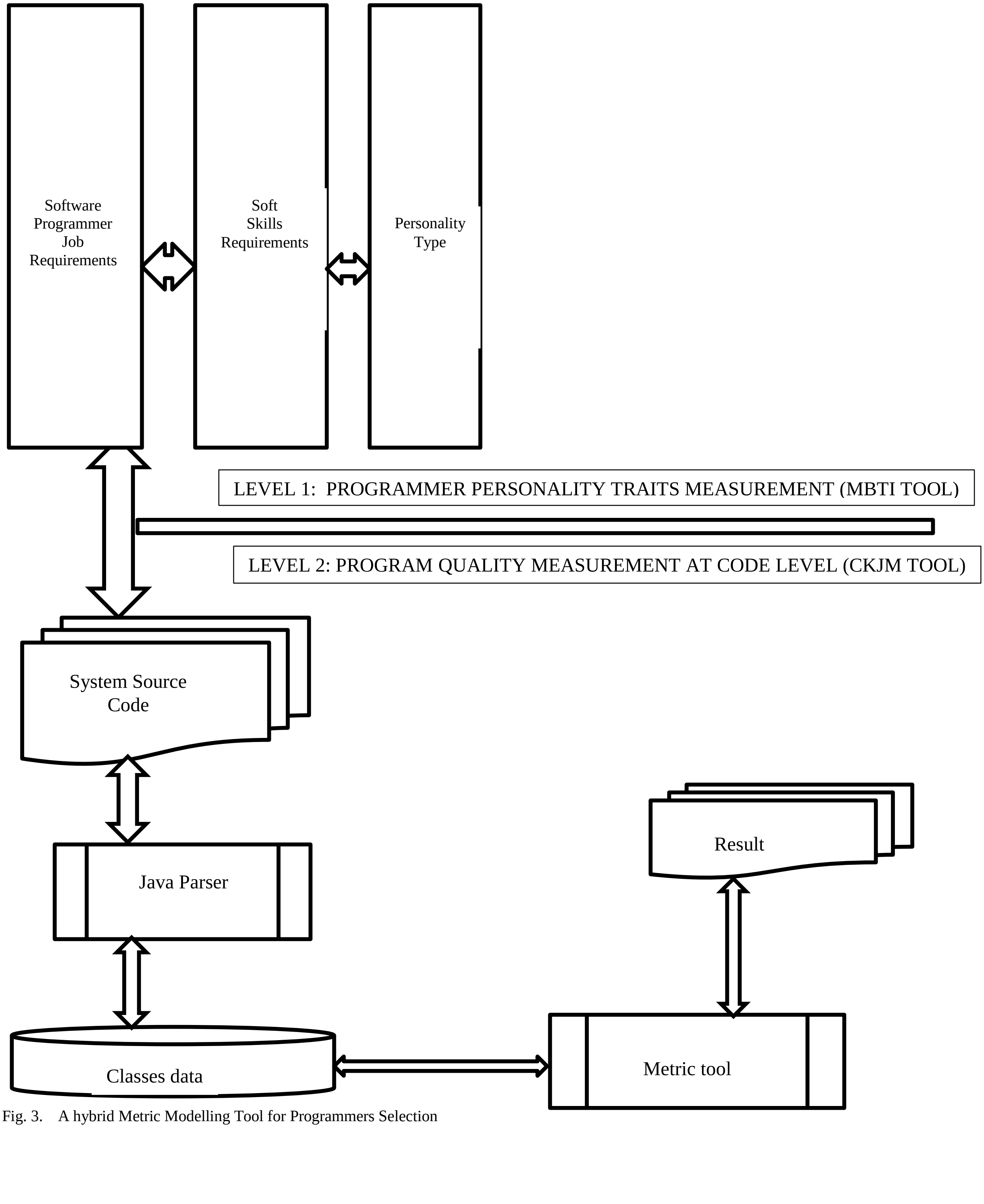

Software Programmer Job Requirements
Soft Skills Requirements
Personality Type
LEVEL 1: PROGRAMMER PERSONALITY TRAITS MEASUREMENT (MBTI TOOL)
LEVEL 2: PROGRAM QUALITY MEASUREMENT AT CODE LEVEL (CKJM TOOL)
System Source Code
Java Parser
Result
Classes data
Metric tool

Fig. 3. A hybrid Metric Modelling Tool for Programmers Selection

III. LITERATURE REVIEW

Code level measurement of program quality have been studied using class cohesion , coupling and other metrics from the Chidamber and Kemerer metric suite [7,8,9,16,17,19]. High cohesion, range [0,1] and low coupling imply good design. The term cohesion is defined as the "intramodular functional relatedness" in software [22]. Chidamber and Kemerer [19] first defined a cohesion measure for objected oriented software- the Lack of cohesion in Methods (LCOM) metric. Okike [7] studied class cohesion measurement in object oriented systems using Chidamber and Kemerer Metric suite and Java as case study. The study involved 6 different types of Java based industrial systems with over 3000 classes. The result of the study showed that the Lack of Cohesion in Methods metric (LCOM) defined by Chidamber and Kemerer was suitable in measuring class cohesion in the studied systems. In addition the study showed that the LCOM metric satisfy measurement theory conditions, and although the metric is prone to outliers; a new metric was defined which normalizes the LCOM metric such that outliers were eliminated. Furthermore, a pedagogical evaluation and discussion about the Lack of Cohesion in Methods metric using field experiments is presented in Okike[16], while a normalized Lack of Cohesion in Methods metric is presented in Okike [17]. In both studies, the usefulness of LCOM metric alongside Coupling between Object (CBO) and Number of Public Methods (NPM) in the evaluation of well-designed classes were clearly established . Hence by measuring cohesion using the LCOM, CBO and NPM metrics in this study, well designed codes by individual programmers were identified

Furthermore, the Myers Briggs Type Indicator(MBTI) has been widely used by researchers to measure the personality traits of individuals in various capacities and dimensions. Okike and Olanrewaju [2] investigated problem solving and decision making skills of 30 student programmers using the MBTI tool. A decision problem representing a programming task was given to the students. The students were expected to produce computer programs which solves the given problem. The MBTI, an automated personality traits questionnaire based tool was administered on the students. The responses from students were automatically analyzed in order to identify the personality traits of each student. The program code or codes written by each students was also analyzed using a Chidamber & Kemerer Java Metric (CKJM) tool, and the results matched with their corresponding MBTI to determine the problem solving and decision making skill of each programmer by looking at the quality of the resulting program code. The study concluded that The result of this study indicates that among the various personality traits, the Introverted Sensing Thinking Judging (ISTJ) appear to have the best problem solving and decision making skill followed by Introverted Intuitive Feeling Judging (INFJ) compared to other personality traits. However, in all, candidates with personality traits such as Introverted Sensing Feeling Perceiving(ISFP), Introverted Intuitive Thinking Perceiving (INTP), Extroverted Intuitive Feeling Perceiving (ENFP), Introverted Sensing Feeling Judging (ISFJ), Extroverted Intuitive Thinking Judging (ENTJ), Extroverted Sensing Feeling Judging (ESFJ), Extroverted Intuitive Feeling Judging (ENFJ), Introverted Sensing Thinking Perceiving (ISTP), and Introverted Intuitive Feeling Perceiving(INFP) are likely to have averagely problem solving and decision making skills while individuals with Extroverted Sensing Feeling Perceiving (ESFP) and Extroverted Sensing Thinking Perceiving (ESTP) traits appear to have poor problem solving and decision making skills.

Okike [10] investigated the role of personality traits in students' achievements in Computing Science. Results from the study suggests that the strongest motivator for a choice of career in the computing sciences is the desire to become a computing professional rather a students inherent temperamental ability (personality traits). Equally, students' achievements in the computing sciences do not depend only on personality traits, motivation for choice of course of study, and reading habits but also on the use of Internet based sources more going to the university library to use book materials available in all areas.

Okike [11] studied the bipolar factor and systems analysis skills of 60 students analysts at the University of Botswana. The study evolved a new approach to construct a type matrix from a personality type frequency matrix. This approach was used to select the best systems analyst based on personality traits factors.

Bentley [4] reviewed personality traits and programmer characteristics and presented some of the traits that can be indicators of success or failure in computer programming. Weinberg [13] explored the psychology of computer programming and noted that there could be variations in individual productivity due to personality type factor. Capretz [5] investigated personality types of software engineers based on the combined Jung and Myers Briggs bipolar. The study suggested that they were more (Introvert Sensing Thinking Judging (ISTJ) software engineers than other types in his data. Chung [15] studied the cognitive abilities in computer programming using 523 form four secondary school students in Hong Kong. Test administered to the students included mathematics, space, symbols, hidden figures and programming ability. Results of this study suggested that performance in mathematics and spatial tests were significant predictors in programming ability. Similarly, Bishop-Clark and Wheeler [14] investigated the Myers-Briggs personality type and its relationship to computer programming. Using 114 students, the study sought to know if college students with certain personality types performed better than others in an introductory programming course. In this study, results suggested that sensing students performed significantly better than intuition students in programming assignments while judging students performed better than perception students on computer programs although the results were not significant statistically.

IV. STUDY METHODOLOGY

A set of Java based programming task was given to some 33 student programmers who could use the Java programming language confidently. A Chidamber and Kemerer Java metric tool (CKJM) [18] was used to analyse the quality of program codes written by each participating programmer. In addition the Myers Briggs Type indicator (MBTI) was used to measure the personality traits of each participating programmer. In this way, a two level metrics based approach was evolved namely:

Level 1: Human metric tool (MBTI)

Level 2: Code level metric tool (CJKM)

The Human metric tool is based on the Myers Briggs Type Indicator tool. Each participating programmer completed and submitted the automated MBTI questionnaire and was subsequently scored by the tool as to the appropriate personality trait.

At level 2, the programmers were given the same programming task, and each of them developed appropriate Java codes. The codes were evaluated automatically by applying the CKJM tool. The CKJM tool calculates for each program class the following six metrics when used in any experiment [18]

- WMC: Weighted methods per class
- DIT : Depth of Inheritance Tree
- NOC: Number of children
- CBO: Coupling between object classes
- RFC: Response for a class
- LCOM: Lack of cohesion in methods
- Ca: Afferent coupling
- NPM: Number of Public Methods for a class

For the purpose of this paper, the LCOM, CBO and NPM metrics are mainly considered in the assessment of code quality . This follows from earlier research as shown in [7,16,17]. High cohesion range [0,1] and low coupling implies good design. Also the number of methods n in a well-designed class should be less than 5 [17:pg22].

V. Result and Discussion

Table 1 below shows the result of the experiment described above in section 3. The Myers Brigg Type Indicator (MBTI) of each programmer and the Lack of Cohesion in Methods (LCOM) metric of program classes are considered together.

TABLE I. Class Design and Programming Ability of Students Using Chidamber and Kemerer Metric Suite and Mbti

COLS/ 1	2	3	4	5	6	7	8	9	10	11
S/N	**Programmer MBTI**	**Lines of Code**	**WMC**	**DIT**	**NOC**	**CBO**	**RFC**	**LCOM**	**CA**	**NPM**
1	ENFJ	97	8	1	0	0	15	1	0	2
2	ENFP	72	2	1	0	0	11	1	0	2
3	ENFP	70	5	1	0	0	13	10	0	2
4	ENTJ	124	7	1	0	0	17	22	0	3
5	ENTJ	84	2	1	0	0	13	1	0	2
6	ENTJ	139	7	1	0	0	16	21	0	7
7	ESFJ	85	2	1	0	0	9	1	0	2
8	ESFP	130	7	1	0	0	8	13	0	4
9	ESTJ	187	19	6	0	4	46	115	0	5
10	ESTJ	121	6	1	0	0	19	15	0	6
11	ESTJ	59	5	1	0	0	7	1	0	2
12	ESTJ	122	18	6	0	4	42	97	0	4
13	ESTJ	34	2	1	0	1	7	1	0	2
14	ESTP	81	6	1	0	0	17	3	0	6
15	INFJ	89	2	1	0	0	17	1	0	2
16	INFJ	85	2	1	0	0	6	1	0	4
17	INFP	90	2	1	0	1	25	1	0	2
18	INFP	38	5	1	0	0	6	6	1	5
19	INTP	24	2	1	0	0	13	1	0	2
20	INTP	127	20	6	0	3	40	142	0	3
21	INTP	111	2	1	0	0	17	1	0	2
22	ISFJ	129	7	1	0	0	20	21	0	6
23	ISFJ	55	2	1	0	0	21	1	0	8
24	ISFP	63	2	1	0	0	12	1	0	2
25	ISTJ	40	2	1	0	0	12	1	0	2
26	ISTJ	63	2	1	0	0	14	1	0	2
27	ISTJ	58	2	1	0	0	12	1	0	2
28	ISTJ	109	4	1	0	1	25	0	0	3
29	ISTJ	56	9	1	0	0	10	16	1	9
30	ISTJ	250	14	1	0	0	6	6	1	5
31	ISTJ	119	2	1	0	1	31	1	0	2
32	ISTP	125	9	1	0	0	20	34	0	3
33	ISTP	57	2	1	0	0	12	1	0	2

Source: Adapted from [21]

Using the MBTI tool, the personality characteristics of the programmers were established as shown in column 2, while the corresponding program quality characteristics of program codes written by the programmers are shown from columns 3-11 as measured by the Chidamber and Kemerer Java Metric (CKJM) tool [18]. A comprehensive discussion about the Chidamber and Kemerer suite of metrics is presented in [19]. Furthermore, a pedagogical evaluation and discussion about the usefulness of Chidamber and Kemerer's metric suite, particularly the Lack of Cohesion in Methods (LCOM) metric is presented in [7, 16, 17] . Using the CKJM tool the following metric were computed for each class or program module written by a programmer: Weighted Method per class (WMC), Depth of Inheritance Tree (DIT), Number of Children (NOC), Coupling Between Object (CBO), Response for a Class (RFC), Lack of Cohesion in Methods (LCOM), Afferent Coupling (CA), and Number of Public Methods (NPM). Details about each of this metric have been discussed in [7,18,19].

A. *Bipolar Factor Characteristics s of Candidates*

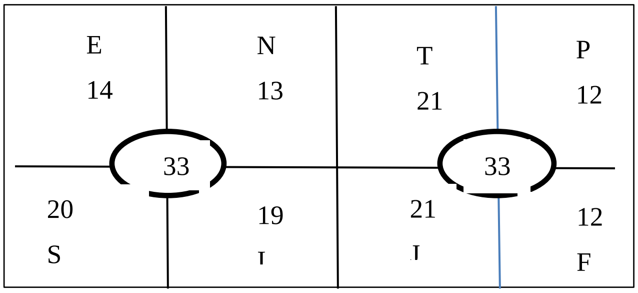
Table 2 below presents the personality frequency matrix of the participating programmers [11].

TABLE II. PERSONALITY FREQUENCY MATRIX

Personality Type		Type Indicator	N
Extroversion(E)	Introversion (I)	EI	3 3
14	19	33	
Sensing (S)	iNtuition (N)	SN	3 3
20	13	33	
Thinking (T)	Feeling (F)	TF	3 3
21	12	33	
Judging (J)	Perceiving (P)	JP	3 3
21	12	33	

From this table, the dominant personality traits are Thinking (T) =21, Judging (J) = 21, Sensing(S) = 20, and iNtuition (N) = 19. Arising from Table 2, a type matrix table is presented in Table 3. Diagonals of type matrix tables must sum up to the total number of participants [11]

TABLE III. TYPE MATRIX TABLE



VI. DISCUSSION

From Table 1, programs with LCOM value in the range [0,1] are cohesive, and hence well designed. These programs were written by candidates with serial numbers 1,2,5,7,11,13,15,16,17,19,21,23-28,31 and 33. The corresponding personality traits of these candidates are ENFJ, ENFP, ENTJ, ESFJ, ESTJ, ESTJ, INFJ, INFJ, INFP, INTP, INTP, ISFJ, ISFP, ISTJ, ISTJ, ISTJ, ISTJ, ISTP (research questions bullet 1 and 2). Overall the number of well designed cohesive programs based on personality traits are shown in Table 4 below

TABLE IV. PERSONALITY TYPE AND GOOD PROGRAM DESIGN

Personality Type	No of Cohesive Programs
ENFJ	1
ENFP	1
ENTJ	1
ESFJ	1
ESTJ	2
INFJ	2
INFP	1
INTP	2
ISFJ	1
ISFP	1
ISTJ	4
ISTP	1

Considering the bipolar factors – Extroversion (E), Introversion (I), Sensing (S), iNtuition (N), Thinking (T), Feeling (F), Judging (J), and Perceiving (P), the number of well designed program codes are shown in Table 5 below

TABLE V. GOOD PROGRAMS BY BIPOLAR FACTOR

Bipolar Factor	No of good programs
Extroversion (E)	10
Introversion (I)	6
Sensing (S)	10
Intuition (N)	8
Thinking (T)	10
Feeling (F)	8
Judging (J)	12
Perceiving (P)	6

Table 5 also provides answers to research questions (bullets 1 and 2) of this study. From this study, introverts appear not have better code design ability than extroverts. In fact, extroverts could be better programmers than introvert (hypothesis bullet 1).

Sensors could design better codes than iNtuitives (hypothesis bullet 2). Thinkers could design better codes than feelers (hypothesis bullet 3). Judges could be better code designers than perceivers (hypothesis bullet 4). The study suggests that there is significant relationship between personality traits and code quality (hypothesis bullet 5). This result is also supported in [2]

VII. Conclusion

In this study, a model for measuring both the personality traits of individual programmers and the quality of programs developed these programmers at two levels has been presented. The model could be used when selecting competent computer programmers since the quality of well designed computer program can be measured by the level of cohesiveness of the program module or class [7,8,9,16,18,19]. In addition, good computer programmers appear to have strong personality traits such as judging, extroversion, sensing, thinking, intuition, feeling, and could have introversion and perceiving abilities. This conclusion supports previous studies a presented in [1, 2,3,20]. Further details about the peculiarities of these traits are fully discussed in [22].

References

[1] F. Capretz, and F. Ahmed, "Making sense of software development and personality types," ITPro, vol. 12, No. 1 January/February 2010.

[2] E. U. Okike and A. Olanrewaju, "Problem solving and decision making: consideration of individual differences in computer programming skills using Myers Briggs Type Indicator (MBTI) and Chidamber and Kemerer Java Metric (CKJM)". *Journal of Applied Information Science and Technology, Vol. 7. No. 1, pp. 27-34 2014* .

[3] D. A. Da Cunha, and D. Greathead, " Does Personality matter? An Analysis of code – review ability," Communications of the ACM. Vol. 50. No. 5, pp. 109-111, May 2007

[4] J. E. Bentley,"Laziness, Impatience, Hubris: Personality Traits of a great Programmer". Unpublished

[5] F. L. Capertz, "Personality types in Software Engineering", *Int. J. Human-Computer Studies.* 58, pp207-214, 2003

[6] R. T. Turley, and J. M. Bieman, "Competencies of exceptional and non exceptional software engineers," J. Systems Software. 28:19-38, 1995

[7] E. U. Okike , " Measuring class cohesion in Object Oriented System Using Chidamber and Kemerer Metric Suite and Java as case Study". Ph.D Thesis Department of Computer Science, University of Ibadan, 2007. Unpublished

[8] L. Badri and M. Badri, "A proposal of a New Class cohesion Criterion: An experimental study" *Journal of Object Technology*, vol. 3, no. 4, pp. 145-159, April 2004

[9] M. Bieman and B. K. Kang , "Cohesion and reuse in Object Oriented Systems", Proceedings of the Sysmposium on Software Reusability (SSR'95), Seatle WA, pp. 259-262, April 1995

[10] Ezekiel U Okike, "Investigating students'achievement in computing science using human metric", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 5. No. 4, pg. 180-186, 2014

[11] E. U. Okike. "Bipolar Factor and Systems Analysis Skills of Student Computing Professionals at University of Botswana, Gaborone," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 5. No. 3, 2014

[12] M. Omar, and S Syed-Abdullah, "Identifying effective software engineering (SE) team personality types composition using rough set approach," IEEE 2010

[13] G. M. Weinberg, "The Psychology of Computer Programming", 2nd ed. Van Nosttrand Reinhold:New York, 1998

[14] C. Bishop-Clark and D. Wheeler, "The Myers-Briggs personality type and its relationship to computer programming," Journal of Research on Computing in Education. Vol. 26 Issue 3, pg. 358-371, 1994

[15] C. Chung, "Correlates of problem solving in programming" CUHK Educational Journal Vol. 16. No. 2 pp185-190, 1986

[16] Ezekiel Okike, "A Pedagogical Evaluation and Discussion about the Lack of Cohesion in Method (LCOM) Metric Using Field Experiment", *IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 3, pg. 36-43,* March 2010

[17] Ezekiel Okike, "A Proposal for Normalized Lack of Cohesion in Method (LCOM) Metric Using Field Experiment", *IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 5, pg. 19-26, July* 2010

[18] S. Diomidis "Tool writing: A forgotten art?". *IEEE Software*, 22(4), pp.9-11, July/August 2005.

[19] S. R. Chidamber and C. F. Kemerer , "Toeards a metric suite for Objected Oriented Design, Object Oriented Programming Systems, Languages and Applications", *Special Issue of SIGPLANNotices, vol. 26, No.6, pg 476-493, 1994*

[20] Tieger, Paul D. and Barbara Barron- Tieger , "Do what you are. Boston: Little, Brow, and company, 2001.

[21] A. Olanrewaju, "Problem solving and decision making: consideration of individual differences in computer programming skills using Myers Briggs Type Indicator (MBTI) and Chidamber and Kemerer Java Metric (CKJM)", M.SC Dissertation, Department of Computer Science, University of Ibadan Nigeria, 2012. Unpublished.

[22] E. Yourdon, and L. Constantine, "Structured design: Fundamentals of a Discipline of computer program and systems Design, Englewood Cliffs, New Jersey: Prentice-Hall, 1979

Segmentation of Acute Lymphoblastic Leukemia Using C-Y Color Space

Reham Mohammed
Dept of Computer Science,
Faculty of Computers and Information
Mansoura University, Egypt

Omima Nomir
Dept of Computer Science,
Faculty of Computers and Information
Mansoura University, Egypt

Iraky Khalifa
Dept of Computer Science,
Faculty of Computers and Information,
Helwan University, Egypt

***Abstract*—Medical image analysis process usually starts with segmentation step, which aims to separate different objects in the image scene. This is achieved by mainly dividing the image into two parts, the region of interest (ROI) and the background. Segmentation of acute lymphoblastic leukemia blood cell (ALL) based on microscope color image is one of the important step in the recognition process. This paper proposed a technique which aims to segment the color image of acute leukemia by transforming the RGB color space to C-Y color space .in the C-Y color space, the luminance component is used to segment (ALL) .The proposed algorithm runs on 100 microscopic ALL images and the experimental result shows that the proposed system can provide a good segmentation of ALL from its complicated background and shows that the segmentation accuracy of the proposed technique is 98.38% compared to the result of the manual segmentation method by expert.**

Keywords—Image Segmentation; acute lymphoblastic leukemia; RGB; C-Y color space

I. Introduction

Cancer is a class of diseases characterized by out-of-control cell growth. There are many different types of cancer, and each is classified by the type of cell that is initially affected, one of them is Leukemia. Leukemia is a cancer that begins in the bone marrow. It is caused by excessive production of leucocytes that replace normal blood cells. There are four major different types of Leukemia according to the growth speed overproduction of leukemic cells [1]. The four main types of Leukemia are : Acute lymphoblastic leukemia (ALL) ,Acute myelogenous leukemia(AML),Chronic lymphocytic leukemia(CLL) and Chronic myelogenous leukemia (CML).This work focuses on the segmentation of Acute lymphoblastic leukemia (ALL), which is called also childhood leukemia.The early and fast identification of the leukemia type, greatly aids in providing the appropriate treatment for the specific type. Over years several works have been conducted in the area of general automatic segmentation and detection methods of ALL. Most of the methods are based on local image information .Cell segmentation using active contour models is presented in [3].

In [4] proposed a blood image segmentation algorithm based on automatic thresholding and binary filtering .in [5] proposed segmentation technique based on watershed transform to extract the nucleus distribution information is used to extract cytoplasm from the background including RBC. While effective for nucleus segmentation this method fails when the cytoplasm is not round. Color images allow more reliable image segmentation and provide a better description of a scene than grayscale images [2] many approaches produced several techniques that using transformation of original RGB images into different color space as in [7] proposed algorithm that converts the RGB color space to HSV color space by working on H channel after extracting it from HSV color space. In [8] it uses Lab color space for segmentation process.

In [9] proposed algorithm Based on HSI color space, enhancement technique .In this paper we proposed a segmentation algorithm on digital microscope images for acute lymphoblastic leukemia based on C-Y color space.

II. Methodology

The main goal of this work is to segment the microscopic image of ALL by converting the RGB color space to C-Y color space after extracting the Y channel from the C-Y and apply median filter on it .we present a comparison of our algorithm's accuracy and the accuracy of RGB segmentation algorithm.

A. *The Dataset*-The images of the database have been captured with an optical laboratory microscope coupled with a Canon Power Shot G5 camera. All images are in JPG format with 24 bit color depth, resolution 2592 × 1944. The images are taken with different magnifications of the microscope ranging from 300 to 500. The ALL-IDB database has two versions [10]; ALL-IDB2 version of the database is used.

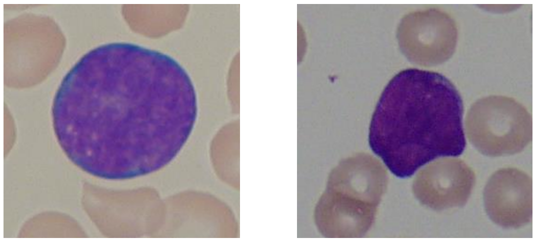


Fig. 1. Example of ALL Dataset

B. *C-Y Color space*

Through reviewing several techniques that used transformation of RGB color space to different other color space, it has been noted that some color space has complex equations of transforming like HSI. So we used the C-Y color space. The C-Y color model has three color components B-Y, R-Y, G-Y, and one luminance component Y [11]. Only two of three color components are needed to define a color.

The conversion of the RGB color space to C-Y color space can be computed using the following 3 × 3 transformation matrix as in (1):

$$\begin{bmatrix} Y \\ R-Y \\ B-y \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.701 & -0.587 & -0.114 \\ -0.299 & -0.587 & 0.886 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad (1)$$

Here, both R-Y and B-Y represent the chromaticity of a color in C-Y color space.

$$S = \sqrt{(R-Y)^2 + (B-Y)^2} \qquad (2)$$

$$\theta = \begin{cases} \tan^{-1}\binom{R-Y}{B-Y} & for\ S \neq 0 \\ undefined & for\ S = 0 \end{cases} \qquad (3)$$

In C-Y color model, saturation (S) as in (2) and hue as in (3), θ can also be derived from the R-Y and B-Y components as above equations.

In order to convert the C-Y color space to RGB color space we use the following transformation matrix in (4):

$$\begin{bmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.1 & 0.0 \\ 0.1 & -0.509 & -0.194 \\ 0.1 & 0.0 & 0.1 \end{bmatrix} \begin{bmatrix} \mathrm{Y} \\ \mathrm{R-Y} \\ \mathrm{B-Y} \end{bmatrix} \qquad (4)$$

C. *Proposed approach*

The essential goal of acute leukemia blood cell segmentation is to extract structural component such as blast from its complicated background by using C-Y color model. The proposed algorithm for ALL image segmentation is given below.

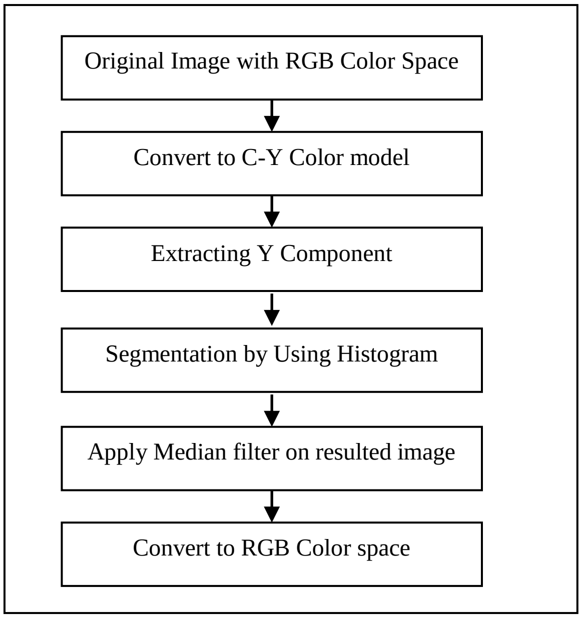

Original Image with RGB Color Space
Convert to C-Y Color model
Extracting Y Component
Segmentation by Using Histogram
Apply Median filter on resulted image
Convert to RGB Color space

Fig. 2. Proposed system for ALL segmentation

The technique for segmentation ALL by C-Y color model composed of the following steps which are summarized below.

Step 1: transforming the original image with RGB color space to C-Y color space using the transformation matrix.

Step 2: Extracting the luminance Y component from C-Y color space.

Step 3: Select the threshold value using Y component from C-Y color space from the histogram.

Step 4: Applying the median filter N × N (N = 5) to the resulted image.

Step 5: Convert the resulted image to RGB & display the result.

III. Experimental Results and Discussion

In this presented work we introduced an approach of segmentation the acute leukemia blood cell (ALL) based on C-Y color space. The result obtained after applying the proposed algorithm as shown below.

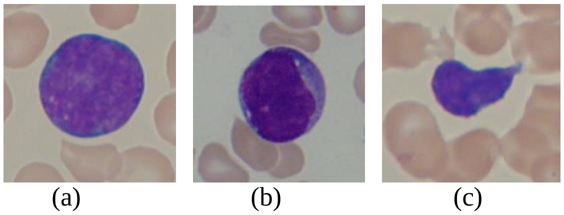

(a) (b) (c)

Fig. 3. are three different examples of original RGB images

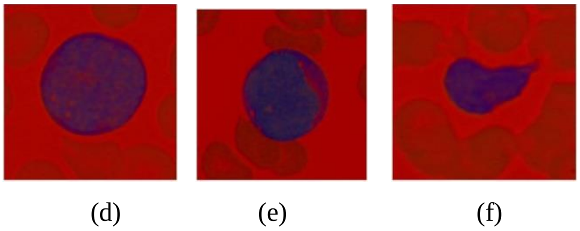

(d) (e) (f)

Fig. 4. (d,e,f) are three Equivalent C-Y Images for a,b,c

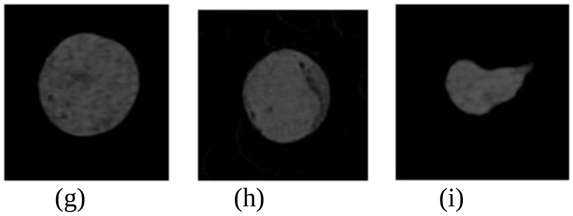

(g) (h) (i)

Fig. 5. (g ,h,i) are three B-Y components of C-Y for d ,e , f

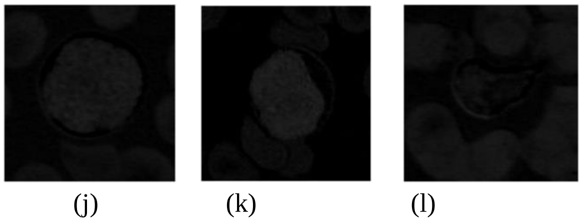

(j) (k) (l)

Fig. 6. (j,k ,l) are three R-Y components of C-Y for d ,e , f

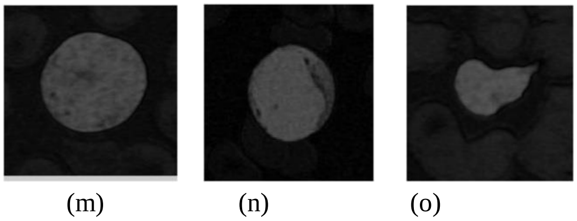


Fig. 7. (m,n,o) are three S components of C-Y for d ,e , f

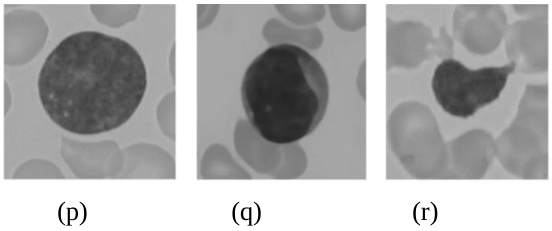


Fig. 8. (p,q,r) are three Y components of C-Y for d ,e , f

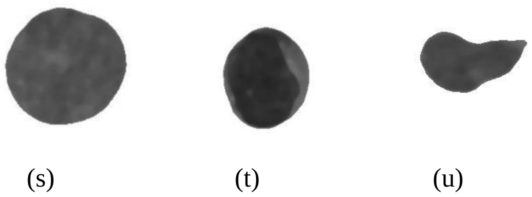


Fig. 9. (s,t,u) are three segmented images of Y component for (p,q,r)

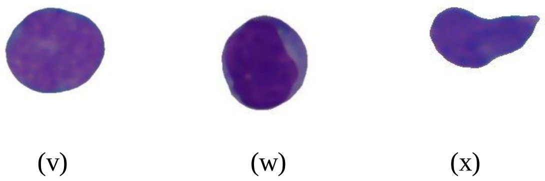


Fig. 10. (v,w,x) are the converted of (s,t,u)images into RGB

The proposed methodology is tested using 100 images of ALL cells and it has successfully isolated the blast part of the cell from the other components of the image this segmentation based on the luminance (Y) component of C-Y color space. To validate the proposed method statistically, the global quantitative method is used. A comparison on area pixels between the accuracy of C-Y algorithm and RGB algorithm according to the expert result was done to measure the accuracy of the method in quantitative manner.

This next table shows a performance comparison of acute leukemia image segmentation accuracy using C-Y and RGB color spaces. For image segmentation using RGB and C-Y color space observation it is found that the method based on RGB color space has not performed well .Means it should not gives accurate result, Besides that ,the shape of the blast after the segmentation process is not quite similar to the expert blasts. While the method based on C-Y color space using Y component can provide a better accuracy segmentation and shape to the expert blasts.

TABLE I. THE RESULTS OF SEGMENTATION'S ACCURACY BY C-Y COLOR SPACE COMPARED WITH THE SEGMENTATION'S ACCURACY BY RGB COLOR SPACE

Type of cell	Segmentation accuracy
C-Y	98.38%
RGB	95.93 %

IV. CONCLUSION

In this paper we proposed a technique of segmentation of ALL microscope images by using C-Y color space. This segmentation is based on the luminance (Y) component of C-Y color space. This approach also showed that applying the C-Y color space is simpler than other color space due to their complex equations and show good result in segmentation of ALL microscopic color images. The segmentation accuracy for the tested images is 98.38 % compared to the accuracy of segmentation of RGB color space 95.93%. The results show that our segmentation technique is robust. In the future, the result of this work can be used as the basis for extracting the other features from the acute lymphoblastic leukemia blood samples.

REFERENCES

[1] The Cancer Institute of New Jersey Patient Education Committee, Leukemia and Lymphoma awareness month , New Jersey , USA , 2008.

[2] Aimi Salihah , A.N, M.Y.Mashor , Nor Hazlyna Harun " Colour Image Enhancement Techniques for Acute Leukemia Blood Cell Morphological Features ", IEEE pp.3677-3682, 2010 .

[3] G. Ongun. U. Halici. K. Leblebicioglu . V. Atalay. M. Beksac , and S. Beksac." An automated Differential Blood Count System" .In Int.Conf. off IEEE Engineering in Medicine and Biology society , Volume 3, pages

[4] Angulo J, Flandrin G. Microscopic image analysis using mathematical morphology: application to haematological cytology, 1:304–312, 2003.

[5] Dorini LB, Minetto R, Leite NJ. White blood cell segmentation using morphological operators and scale-space analysis. Proc Brazilian Symp Comput Graph Image Proc, 294–304, 2007.

[6] M. Veluchamy, K. Perumal and T. Ponuchamy, "Feature Extraction and Classification of Blood Cells Using Artificial Neural Network," American Journal of Applied Science, vol. 9, pp. 615-619, 2012.

[7] N.Sinha and A. G. Ramakrishman. Automation of Differential Blood Count. In Proceeding Conference on Convergent technologies for Asia - pacific Region .2:547-551.2003.

[8] S. Mohapatra and D. Patra, "Automated Cell Nucleus Segmentation and Acute Leukemia Detection in Blood Microscopic Images," in International Conference on Systems in Medecine and Biology, India, 2010.

[9] N. H. A. Halim, M. Y. Mashor, A. S. Abdul Nasir, N. R. Mokhtar, and H. Rosline. 2011. Nucleus segmentation technique for acute leukemia, in Proceedings of the IEEE 7th International Colloquium on Signal Processing and Its Applications (CSPA '11), pp. 192–197, March 2011.

[10] Donida Labati, R., Piuri, V., Scotti, F. ALL-IDB:the Acute Lymphoblastic Leukemia Image DataBase for image processing, 2011.

[11] E. R. Dougherty, Electronic Imaging Technology, SPIE-The International Society for Optical Engineering, 1999. 2583- 2586.2001.

Clustering of Slow Learners Behavior for Discovery of Optimal Patterns of Learning

Thakaa Z. Mohammad
Department of Computer Science,
Faculty of Computer and Information Sciences,
Ain Shams University,Cairo,Egypt

Abeer M.Mahmoud
Department of Computer Science,
Faculty of Computer and Information Sciences,
Ain Shams University,Cairo,Egypt

***Abstract*—with the increased rates of the slow learners (SL) enrolled in schools nowadays; the schools realized that the traditional academic curriculum is inadequate. Some schools have developed a special curricula that are particularly suited a slow learner while others are focusing their efforts on the devising of better and more effective methods and techniques in teaching. In the other hand, knowledge discovery and data mining techniques certainly can help to understand more about these students and their educational behaviors. This paper discusses the clustering of elementary school slow learner students behavior for the discovery of optimal learning patterns that enhance their learning capabilities. The development stages of an integrated E-Learning and mining system are briefed. The results show that after applying the clustering algorithms Expectation maximization and K-Mean on the slow learner's data, a reduced set of five optimal patterns list (RSWG, RWSG, RWGS, GRSW, and SGWR) is reached. Actually, the students followed these five patterns reached grads higher than 75%. Therefore, the proposed system is significant for slow learners, teachers and schools.**

Keywords—Data mining; E-learning; Slow Learners

I. Introduction

A child may be a slow learner for various reasons, including: heredity, inadequate brain development due to lack of stimulation, low motivation, attention problems, behavior problems, different cultural background from that which dominates in the school, or distracting personal problems[1-4]. With so many slow learners remaining in school nowadays, the schools are beginning to realize that they make special provision for such students. The traditional academic curriculum of past years is now judged inadequate. Some have developed special curricula that are particularly suited a slow learners and their needs. Some schools, rather than developing special curricula are focusing their attention on the devising of better and more effective methods and techniques for use in teaching the regular curricula to slow learners [5]. Other schools work on both directions [6].

Knowledge discovery is an evolving interdisciplinary field that is connected to a number of research areas containing intelligent and adaptive web-based educational systems, intelligent tutoring systems, adaptive hypermedia, online courses mining systems and more others [7, 8]. Recently, its applications especially in the educational E-learning environments have been improved in order to cluster and mine the characteristics and the records of the learners to predict their studying results [9-11]. Also, it can find out helpful information that can be utilized informative estimation to aid educators to establish an educational basis for decisions when modifying and designing approach or teaching environment. Actually, in the educational systems, the data mining application is a repeated cycle of testing, refinement and hypothesis formation [7]. Therefore, the educational knowledge discovery and data mining techniques certainly can help through the discovery of hidden valuable knowledge to understand more about slow learner students and their educational behaviors.

This paper discusses the discovery of the optimal pattern of learning for elementary school slow learner students through applying two machine learning clustering algorithms Expectation maximization and K-Mean. The development stages of the proposed integrated E-Learning and mining system are briefed, where it goes through the development of digital contents of the English course with a mapping of these digital materials with environmental background of these students. Then the Skelton of the database with a simplified graphical user interface suitable for these slow learners is briefly presented. The rest of the paper is organized as follow; section 2 presents literature review and related. Section 3; go through the slow learner students definition, characteristics, suitable strategies of teaching and recommendation. Traditional teaching versus E-learning teaching is in Section 4. Section 5, presents machine learning clustering algorithms. Section 6 discusses the proposed clustering based integrated system based on three main levels, these are bottom, middle and top. Section 7, concludes the paper.

II. Literature Related Work

Data mining is supported by hosting models or tasks that capture the characteristics of data in several different ways such as: classification, clustering and visualization and other models. Many studies tried to enhance the results of implementing data mining approaches through e-learning systems [7-11]. But still there is an urgent need for considering the learning environment's design in order to utilize the chance provided by the internet. E-environment is not just a conversion from print-based material to digital one.

TABLE I. LITERATURES OF DIFFERENT DATA MINING TASKS IN EDUCATIONAL E-LEARNING

<table>
<tr><th colspan="3">Classification</th></tr>
<tr><th><em>Author</em></th><th><em>Objective</em></th><th><em>Results</em></th></tr>
<tr><td>Kangaiammal et. al (2013)</td><td>Classifying the user learning activities during the learning process depending on a continuous evaluation test for recognizing the understanding level.</td><td>Using a Rough Set Approach, they could increase teacher ability of awareness of the user learning ability before preparing the content of the course.</td></tr>
<tr><td>Marijana et. al (2009)</td><td>Creating adaptive courses for e-learning depend on the style of learning utilizing the intelligence tools</td><td>The students accomplished good results & high satisfaction's while attended the adapted courses based on learning styles.</td></tr>
<tr><td>Aski & Torshizi (2009)</td><td>To compare and investigate the results of the perceptions of four classification tools to analyze and classify the information of learner.</td><td>It is discovered that tools which use the Simple Bayesian or Decision Tree Algorithms had more truthful outcomes and helpful means in classifying the learner's information.</td></tr>
<tr><td>Furkan (2008)</td><td>To classify examination performances of three intelligence artificial favorite tools ANFIS, SVM, and ANN among environment that depends on E-Learning)</td><td>The system of Adaptive-Network-Based Fuzzy Inference (ANFIS) achieved better performances than Support Vector Machine (SVM) and Artificial Neural Network (ANN).</td></tr>
<tr><th colspan="3">Clustering</th></tr>
<tr><td>Mamcencko et. al (2011)</td><td>To analyze the data of the electronic examination.</td><td>The association rules and clustering aid in defining the relationship and patterns in the data of electronic exam. Also enhanced the system of E-examination by descriptive model.</td></tr>
<tr><td>Anitha & Krishnan (2011)</td><td>Build a model to link the E-learners at their early stages of learning by presenting navigation recommendation</td><td>They combined the clustering task with (AR) technique to achieve their goals. Their results showed that the usage of the patterns of clustered access decreased the size of data set and enhanced the accuracy of recommendation</td></tr>
<tr><td>Dominguez, et. al (2010)</td><td>To presents a method where the student current & past data is utilized live to produce hints for students that are ending the exercises of programming during the online competition</td><td>Association's rules- clustering and numerical analysis helped them discovering that the users who are given hints achieved higher marks than the users who were not.</td></tr>
<tr><td>Carmona et al. (2010)</td><td>To present the subgroup discovery techniques' application to the E- learning data from the Learning Management System (LMS) of the universities of Andalusia.</td><td>Optimization-Evolutionary algorithms were used for reducing a group of comprehensible rules that were obtained (due to their usage of the linguistic labels as well as their tiny size) that make them more explanatory for the instructor as well as getting the same values in the other measures of quality.</td></tr>
<tr><td>Liu (2009)</td><td>To generate the characteristics of learners from his data.</td><td>Achieved the behavior description of learner throughout dynamically generated metric and measurements.</td></tr>
<tr><th colspan="3">Visualization</th></tr>
<tr><td>Prema & Prakasam (2013)</td><td>Increasing the quality of the content in the learning materials as well as enhancing the concepts of self –learning for the students, and increasing their examination performance</td><td>It is indicated that there is a positive results for the usage of Data mining based e-learning system on the quality of learning and teaching.</td></tr>
<tr><td>Hung & Saba (2012)</td><td>To suggest a generic model for Educational Data Ming (EDM) examined by the existed model of the data mining and EDM literature.</td><td>The case study displayed the relationships & patterns that are exposed from the model of EDM that could be applied. The specific mining techniques of education help in improving pedagogical decision making and instructional design.</td></tr>
<tr><td>Kazanidis Et. al, (2009)</td><td>To suggest a platform that depends on the framework for recording, analyzing and processing data from Learning Management Systems (LMS)</td><td>The benefits of the usage of the frame work utilized the tools of data mining (DM) for the evaluation of the users and the content, suggesting new metrics and indexes to be utilized with the algorithms of DM, and to be adaptable to any LMS.</td></tr>
</table>

Table 1 abstracts some of these tries categorized by different mining task. These papers focuses generally on educational data mining model with different categories but certainly, they helped us reaching the results of this paper. Actually, Table 1, innovated the idea of our study which cover the effects of e-learning systems on the slow learners.

III. SL: CHARACTERISTICS & RECOMMENDATIONS

A child may be a slow learner for various reasons, including: heredity, inadequate brain development due to lack of stimulation, low motivation, attention problems, behavior problems, different cultural background from that which dominates in the school, or distracting personal problems[1-4]. The slow learning is not a learning disability or diagnostic category. It means that these students suffer from low rate of understanding for the materials; therefore, they need special education strategies. The slow learner is those students that when you are setting up the lesson, they cannot find his or her materials (book, pencils, Papers), when you remind him about the last lesson, he or she doesn't seem to remember anything. Table 2, list some of slow learners characteristics, teaching strategies and recommendation.

TABLE II. ABSTRACTION OF SL CHARACTERISTICS, TEACHING STRATEGIES AND RECOMMENDATION

Characteristics [17,28]
• Scores low rates on evaluation tests • Their classification below average ability • Functioning ability is below grade level • prefers playing with younger children • Faces difficulty in following multi-step directions • Frequently has impaired fine motor coordination such as delayed ability to tie shoe laces • Has few internal strategies (i.e. organizational skills, transferring/generalizing information) • Works well with "hands-on" material (i.e. labs, pictured texts, manipulative, activities) • May have poor self-image & lacks self-confidence • Works on all tasks slowly • Masters skills slowly or does not master at all
Teaching Strategies [1-4, 17,28,29]
• Inclusion of students with slow learner into regular classes is generally an effective strategy • Another strategy is to enroll the child in the least demanding syllabus available and supplement classroom learning with one-on-one teaching by special educators and occupational therapists • The strategy to teach slow learner by E-learning environment may enhance their educational behavior
Recommendation [1-4, 29]
• Slow learner should receive special help outside the classroom. • Teacher should spent great deal of time with slow learner. • Do not give slow learner any designation which indicates they are in fact "slow learners." • Look for every opportunity to encourage and to reinforce the idea that the students are improving • Use tighter lesson plan because slow learners cannot usually think very creatively or spontaneously. • Prepare a lesson's content that is concrete, visual, familiar, and personally interesting to the students • Use eye-catching materials such as colored chalk or magic marker for key words in the lesson. • Keep information to no more than five pieces at one time? Because they suffer from short memory and short memory can hold 5-9 items only at a time

TABLE III. BENEFITS & DRAWBACKS OF TRADITIONAL LEARNING VERSUS E-LEARNING

	Traditional Learning	E-Learning
Student Contribution	• A single student in a given period of time can express herself or himself. • just one topic which cans the student expressing his opinion at a time	• Multiple students can share their contributors in the debates and express their opinions. • Multiple topics can be convened simultaneously and the student can be involved concurrently in various topics and express his/her opinion • more ability to understand complex concepts for multiple students
Teacher Contribution	• The teacher commonly talks and speeches more than the learner.	• the learner talks more than or equal to the teacher.
Interaction between Students	• less interaction and is not recorded in reports	• higher interaction and it can be reflected in the amount of the messages which are transferred between the learners in the groups of the study and in the reports of students
Student Control of Learning Process	• can't control the process of learning	• can control the process of learning and to log in into any course at any convenient times & when they feel that they able to receive information
Student freedom	• limited freedom	• Has freedom of the social restraints of gender, perspiration and the appearance
Student Motivation	• Low motivation	• High motivation because of technology
limits of learning	• limited to place and time	• no limit on place or time
Discussion Style	• is done by the whole class	• Happens in group styles or individually.
Teacher role	• The teacher role is the authority,	• The teacher role is to direct the learners into the information and the knowledge
Subject Matter	• the teacher precedes the lesson in accordance with the curriculum and the study program	• the process of determinate the subjects, the study depends on different sources of information, such as net-experts and data banks which are located by the learner.
Learning Emphases	• the student learn "what" only because the teacher is usually busy with finishing the required topics	• the student learn "what" and "how", because , the process of learning contains research study that merges searching for and gathering information from the data banks.

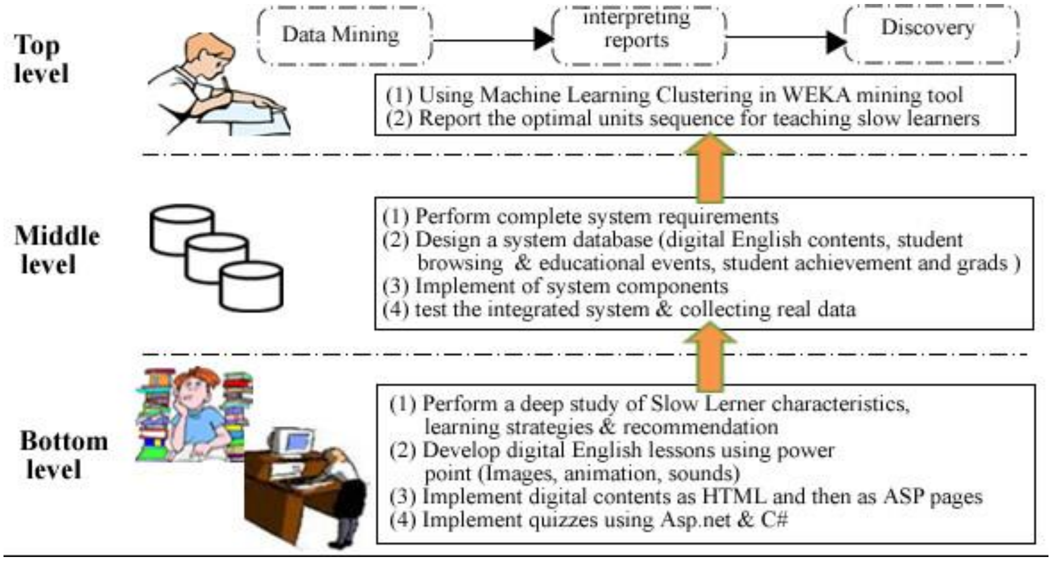

Top level
Data Mining
interpreting reports
Discovery
(1) Using Machine Learning Clustering in WEKA mining tool
(2) Report the optimal units sequence for teaching slow learners
Middle level
(1) Perform complete system requirements
(2) Design a system database (digital English contents, student browsing & educational events, student achievement and grads)
(3) Implement of system components
(4) test the integrated system & collecting real data
Bottom level
(1) Perform a deep study of Slow Lerner characteristics, learning strategies & recommendation
(2) Develop digital English lessons using power point (Images, animation, sounds)
(3) Implement digital contents as HTML and then as ASP pages
(4) Implement quizzes using Asp.net & C#

Fig. 1. The proposed integrated system architecture

IV. Traditional Learning Versus E-Learning

The process of learning can be defined as providing the knowledge for the learners by various ways such as, questioning, doing, watching and listening. Traditional learning is also known as customary education, conventional education or back-to-basics. It refers to the traditional customs which are found for a long time in the schools and the society deemed appropriate traditionally [30]. In the other hand, E-Learning is using the network and computer in teaching and transferring the knowledge and skills. These applications and processes contain “computer based learning”, “Web based learning”, digital collaboration and virtual classrooms. The content of the learning is transferred via internet/extranet, CD-ROM, video tape, audio tape, Internet, satellite TV. E-learning includes media in text form, streaming video form, image, animation and audio form, and it can be instructive or self-paced [30]. Table 3, shows the traditional learning versus e-learning benefits and drawbacks.

V. Clustering

Clustering is unsupervised learning that group the objects into subsets of similar features, where it can be used in many fields, including machine learning, data mining, pattern recognition, image analysis and other. Usually, the first step in clustering is to state a mathematical description of similarity [32]. There are number of criteria’s available to measure the similarity between objects where the default measure is Euclidian Distance. Clustering techniques are able to deal with noisy and high dimensional data. In general, the major clustering algorithms can be classified into following categories. (1) *Partitioning algorithm*: given a database of n data records, it constructs k partitions where each partition represents a cluster and k≤n. (2) *Hierarchical algorithm*: a hierarchical decomposition of data objects is created. (3) *Density based algorithm*: its idea is to keep growing a specific cluster as long as the density in the neighborhood exceeds some threshold (4) *Grid based algorithm:* it quantizes the object space into a finite number of cells that construct a grid structure and all of its operations are performed on grid structure. (5) *Model based algorithm*: it hypothesize a model for each cluster and best fit data that model [32-35].

The K-means clustering algorithm is a partition algorithm that performs one level partition of the data records where it first choose k (number of clusters desired) initial centroids. Each point is then assigned to the closest centroid or cluster, and the cluster is then updated based on the updated points assigned to it. The basic steps of k-means clustering are first determine the centroids coordinates then determine the distance of each object to the centroids, gather the objects based on minimum distance and lastly update the centroids[33,34].

VI. the Proposed Clustering Based Integrated System

The proposed clustering integrated system went through three main levels. These are bottom, middle and top, see Figure 1. In the following, a brief presentation for each of them is given.

A. Bottom Level

In this level, all lessons were developed from scratch where they contain image, sound and emotions that together form an attractive lesson environment and also consider the simulation of the slow learner student’s way of thinking and connectivity between various contents. Note that no specific order of lessons is pre-stated on the slow learner’s students. Ex: some students start with reading unit while others start with listening unit and so on. (Microsoft power point slides plus “iSpring” are used by this stage). For every lesson, a quiz was developed for reporting the understanding degree of each student. Additionally, a final exam is also maintained (ASP.net plus C#

TABLE IV. 24 PERMUTATION LEARNING PATTERN ACTIVITIES (R: READING, G: GRAMMAR, W: WRITING, S: SPEAKING)

1-RWGS	7-GSRW	13-WSGR	19-SGRW
2-RWSG	8-GSWR	14-WSRG	20-SGWR
3-RSWG	9-GWRS	15-WRSG	21-SWRG
4-RSGW	10-GWSR	16-WRGS	22-SWGR
5-RGWS	11-GRSW	17-WGSR	23-SRGW
6-RGSW	12-GRWS	18-WGRS	24-SRWG

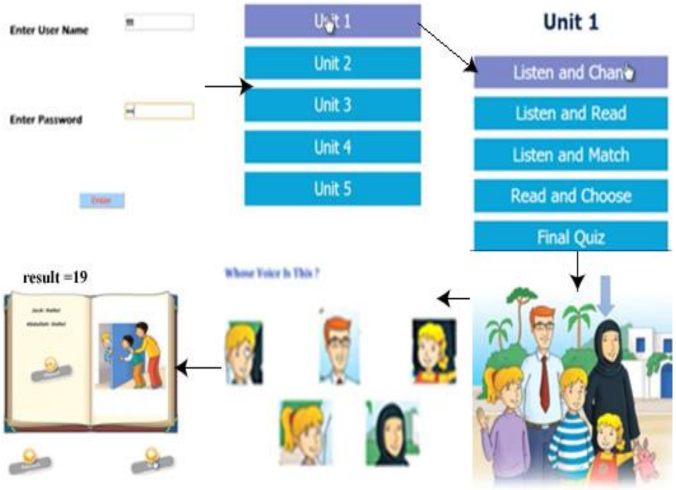
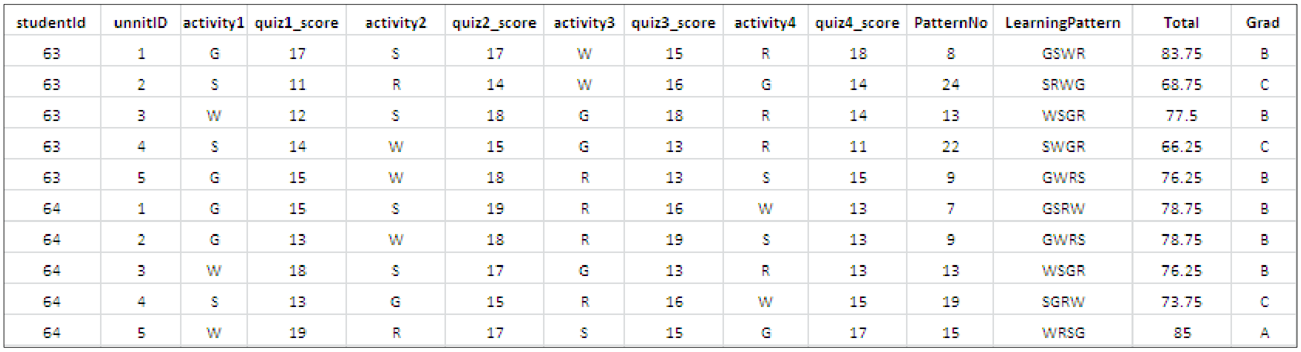


Fig. 2. Ex of a simple GUI during learning



studentId	unnitID	activity1	quiz1_score	activity2	quiz2_score	activity3	quiz3_score	activity4	quiz4_score	PatternNo	LearningPattern	Total	Grad
63	1	G	17	S	17	W	15	R	18	8	GSWR	83.75	B
63	2	S	11	R	14	W	16	G	14	24	SRWG	68.75	C
63	3	W	12	S	18	G	18	R	14	13	WSGR	77.5	B
63	4	S	14	W	15	G	13	R	11	22	SWGR	66.25	C
63	5	G	15	W	18	R	13	S	15	9	GWRS	76.25	B
64	1	G	15	S	19	R	16	W	13	7	GSRW	78.75	B
64	2	G	13	W	18	R	19	S	13	9	GWRS	78.75	B
64	3	W	18	S	17	G	13	R	13	13	WSGR	76.25	B
64	4	S	13	G	15	R	16	W	15	19	SGRW	73.75	C
64	5	W	19	R	17	S	15	G	17	15	WRSG	85	A

Fig. 3. Sample of integrating the whole system collected data

are used to accomplish this stage). Figure 2 shows a sample screen shoot of the developed lessons work flow , where the slow learner first choose the unit no, then follow a specific sequence or learning pattern, to study and perform quizzes and assigments.

B. Middle Level

In this level, the Skelton of the system database and internal structure are achieved. The design of database includes the English course dataset, the student browsing educational events and the student achievement. The SQL server 2008 was used to implement the database ADO.net and is also used to link ASP.net with the database. We have tried to aggregate the sequence log data of the slow learner into a list of features that could capture most aspects of a student's online behavior. The features we have selected are: the login frequency, the date of last login, the time spent online, the number of lessons read, the number of quizzes, the average grade obtained in the unit as a whole, the, the average best grade obtained, and the number of answers to existing questions. All of our features are normalized.

The English course dataset contains four different activities; these are Reading (R), Writing (W), Grammar (G), and Speaking (S). Each slow learner can start his learning using any sequence of these activities, and then continue studding for the rest of activities. Table: 4 shows the possible student learning activities sequences, which various among 24 patterns. Figure 3, show a sample of the slow learner data.

C. Top Level

In this level, the whole consistency and integrity of system components are also achieved. A “WEKA” mining tool is used with target of clustering task. The “WEKA” data analysis tool is a group of machine learning algorithms to solve problems used in real world. Java is the used programming language in “WEKA”. It is freely available software. It is portable & platform independent because it is fully implemented in Java programming language and thus runs on almost any modern computing platform [31]. Weka also contains tools for data preprocessing. In this paper two clustering algorithms is selected from WEKA for interpreting the following results. These are Simple K-means and Expectation Maximization algorithms. Our aim with this analysis was be to determine if selecting a specific learning activities pattern or sequence will affect the overall achievements of the slow learner’s students and affect their educational performance. In addition whether the clusters show mostly qualitative or quantitative differences between the slow learner students or not?.

VII. Experimental Results & Disscussion

Our system is implemented using ASP.NET and C# with SQL database language. The results of this paper are based on random sample of data of the slow learners of elementary school in Kuwait. Each participant accessed the integrated system will participate as a total of five times instead of one time, since the course consist of five units and each unit has a different material. Thus a higher reliability of the results is obtained. Also, the system allows repeating the material to the slow learners in order to enhance their academic achievement based on scientific recommendation for teaching slow learners.

The data used in this paper was collected, and pre-processed based on real slow learner students' information in an elementary school at Kuwait. Actually, the total number of participants was 300 students with a five times login participation for each teaching unit. In each experiment, 500 records were randomly selected.

***Experiment 1**: learning pattern Visualization*

This experiment visualizes the distribution of the slow learner relative to the followed pattern of learning in Figure 4. From the figure, it is obvious that some pattern was followed frequently than other patterns by the students which reflect their interest in such pattern. These are three main patterns (RWSG with frequencies =138, RWGS with frequencies =52 and RSWG with frequencies = 56)

***Experiment 2**: Expectation Maximization algorithm (EM)*
Expectation maximization algorithm is an iterative algorithm for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models. The EM[34] iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. The result of the cluster analysis is as values that indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc. The class indices are sorted according to the prior probability associated with cluster, i.e. a class index of '0' refers to the cluster with the highest probability [34]. Figure 5 shows result of EM algorithm on the selected slow learner data. Figure 6: shows a visualization of results cluster. From both figure, four main clusters were detected from the data and the mapping between the learning patterns relatively to the achievements grads were resulted in the following

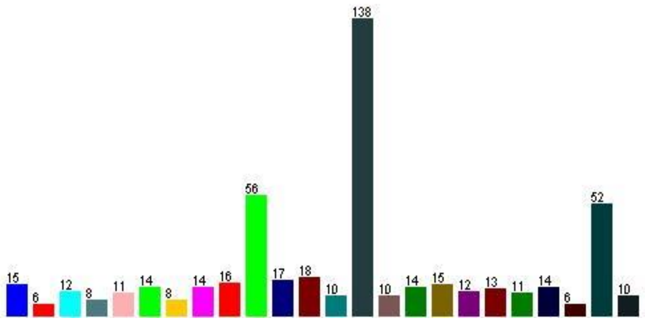


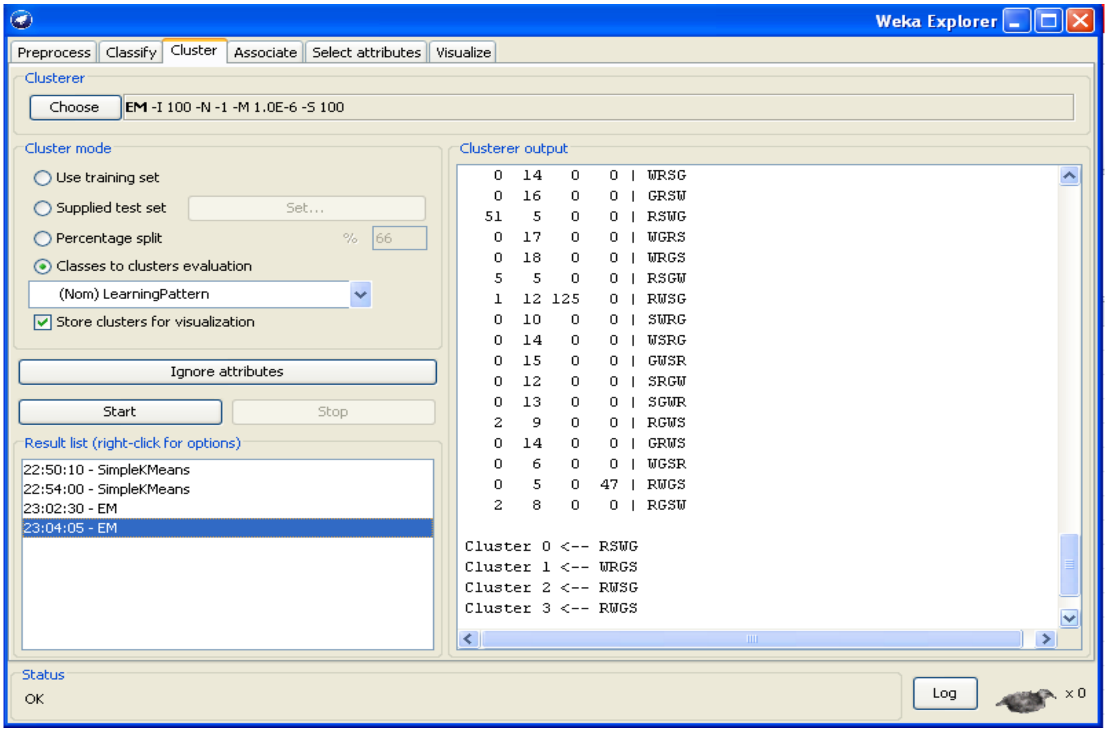
Fig. 4. 24 learning patters with total no of followed students

Cluster 0 ← B ←RSWG
Cluster 1 ← C ←WRGS
Cluster 2 ← D ← RWSG
Cluster 3 ← A ←RWGS

***Experiment 3**: K-Mean cluster with k=2*

K-means clustering algorithm [34] aims to partition n records into k clusters in which each observation belongs to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithms where its main idea is to define k centroids, one for each cluster and a very good recommendation state that these centroids should be placed far away from each other. Applying the K-mean on the slow learner data with two main clusters resulted in the following mapping between the learning patterns relatively and the achievements grads of students.

Cluster 0 ← B ← RWSG
Cluster 1 ← C ← RWGS



Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100

○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) LearningPattern
☑ Store clusters for visualization

22:50:10 - SimpleKMeans
22:54:00 - SimpleKMeans
23:02:30 - EM
23:04:05 - EM

```
  0  14   0   0 | WRSG
  0  16   0   0 | GRSW
 51   5   0   0 | RSWG
  0  17   0   0 | WGRS
  0  18   0   0 | WRGS
  5   5   0   0 | RSGW
  1  12 125   0 | RWSG
  0  10   0   0 | SWRG
  0  14   0   0 | WSRG
  0  15   0   0 | GWSR
  0  12   0   0 | SRGW
  0  13   0   0 | SGWR
  2   9   0   0 | RGWS
  0  14   0   0 | GRWS
  0   6   0   0 | WGSR
  0   5   0  47 | RWGS
  2   8   0   0 | RGSW

Cluster 0 <-- RSWG
Cluster 1 <-- WRGS
Cluster 2 <-- RWSG
Cluster 3 <-- RWGS
```

Status: OK

Fig. 5. Expectation Maximization algorithm results on slow learner data , no of cluster=4

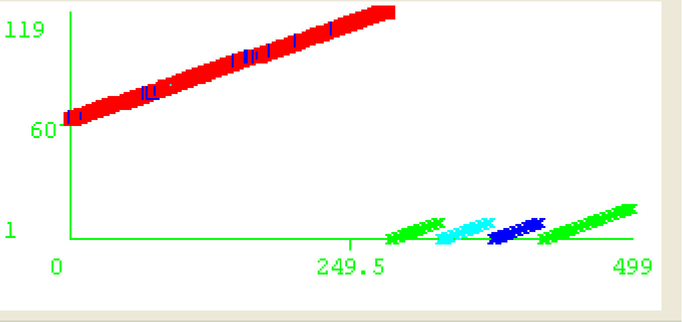


Fig. 6. visualization of Expectation Maximization algorithm on slow learner data, no of cluster=4

Experiment 4*: K-Mean cluster with k=4*

The student's data were only quantitatively differentiated by sequence of learning activities. Applying the K-mean on the slow learner data with four main clusters resulted in the following results

Cluster 0 ← 1 ← 2 ← 3
C ← A ← D ← B
GRSW ← SGWR ← WRGS ← RWSG

Based on results of the three clustering algorithms on the collected slow learner data, an illustration of the relations between patterns and academic achievement is analyzed and hence the three algorithms reached a specific list of optimal learning pattern. This list reduced the 24 proposed learning patterns to a set of 5 optimal patterns list **(RSWG, RWSG, RWGS, GRSW, and SGWR).** Of course, the preferred list from experiment 1 is certainly included in the optimal patterns list. These five optimal patterns helped the slow learner to get grads between A, B or C which are still considered an achievement for the slow learner. In addition, the student who followed pattern RWSG in the three algorithms reached grad B with score higher than 75%. Also, student who followed patterns SGWR or SGWR got an A which is the highest academic achievements for the slow learner. Therefore, these are the optimal learning pattern in learning slow learner student

VIII. Conclusion

With the increased rates of the slow learners enrolled in schools nowadays, the schools realized that the traditional academic curriculum is inadequate. Some schools have developed a special curricula that are particularly suited a slow learner while others are focusing their efforts on the devising of better and more effective methods and techniques in teaching. In the other hand, knowledge discovery and data mining techniques certainly can help through the discovery of hidden valuable knowledge to understand more about these students and their educational behaviors. This paper discussed the discovery of the optimal pattern of learning for elementary school slow learner students through applying two machine learning clustering algorithms Expectation maximization and K-Mean. The development stages of the proposed integrated E-Learning and mining system were briefed, where it goes through the development of digital contents of the English course with a mapping of these digital materials with environmental background of these students.

Then the Skelton of the database with a simplified graphical user interface suitable for these slow learners was also briefed. Based on the results of the three applied clustering algorithms on the slow learner data, a five optimal learning patterns list were concluded as a reduction of 24 possible combination and relative the prepared teaching material. This optimal list includes **(RSWG, RWSG, RWGS, GRSW, and SGWR)** interested patterns and certainly helped the slow learner to get grads between A, B or C which are still considered an achievement for the slow learner. Actually, the student who followed pattern RWSG in the three algorithms reached grad B with score higher than 75%. Also, student who followed patterns SGWR or SGWR got an A which is the highest academic achievements for the slow learner. Therefore, the proposed integrated clustering system is a promising assistant methodology for teaching the slow learner student.

Our suture work includes more investigation for different course material specially in mathematics. Also, different slow learner students with different grads or age level will be studied.

References

[1] http://www.nasponline.org/publications/cq285slowlearn.html

[2] http://www.clubtheo.com/momdad/html/dlslow.html

[3] http://www.foundationosa.org/slow.htm

[4] http://www.gnb.ca/0000/publications/ss/disability.pdf

[5] Lottir Phillps, “A Study of the Preparation of English Teachers for Teaching of Slow Learners”, Olivet Nazarane College, 1970.

[6] Rashmi Rekha Borah,” Slow Learners: Role of Teachers and Guardians in Honing their Hidden Skills”, Int. J. of Educational Planning & Administration, vol. 3, no. 2, pp. 139-143, 2013.

[7] Clark, R,C.& Mayer, R.M.,” E-Learning and the science of instruction”, 3rd edition, san Francisco, CA:Pfeiffer., 2011.

[8] J.willems, “Using Learning styles data to inform e-learning design : A study comparing undergraduates, pstgraduates and e-educators”, Australasian J.of. educational technology, vol.27, no.6, pp.863-880,2011.

[9] Movafegh Ghadirli, H. and Rastgarpour, M.,, “A Model for an Intelligent and Adaptive Tutor based on Web by Jackson’s Learning Styles Profiler and Expert Systems”, Proc.of the Int. MultiConferece of Engineers and Computer Scientists(IMECS 2012) ,vol 1, 2012.

[10] Al-Khalifa, H. S. and Al-Wabel, A. S. 2005. Aided technological methods for special learning: exploratory study. Available online: http://www.gulfkids.com/pdf/Areej.pdf , 2013.

[11] Anitha, A and Krishnan,N. “A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining'. International Journal of Computer ApplicationS, vol.12, no.11,pp.36-41, 2011.

[12] B. M. Ramageri, “Data mining techniques and applications”, J. of CS and Engineering, vol. 1, no. 4, pp. 301-305, 2010

[13] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, 2nd ED, Morgan Kaufmann Publishers, 2006.

[14] P. Berkhin, “Survey of Clustering Data Mining Techniques”, 2002, retrieved 24th Sep, 2012 <http://csis.pace.edu/~ctappert/dps/d861-12/session3-p2.pdf>

[15] Kumar, V. and Chadha, A. “An Empirical Study of the Applications of Data Mining Techniques in Higher Education”, Int.J. of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84, 2011.

[16] Rashmi Rekha Borah, “ Slow Learners: Role of Teachers and Guardians in Honing their Hidden Skills”, Int. J. of Educational Planning & Administration, vol. 3, no. 2, pp. 139-143, 2013

[17] Kangaiammal, A., Silambannan, R., Senthamarai, C., and Srinath, M.,” Student Learning Ability Assessment using Rough Set and Data Mining Approaches’. I.J.Modern Education and Computer Science, vol. 5, pp.1-11, 2013.

[18] Prema, M and Prakasam, S., “ Effectiveness of Data Mining – based E-learning system (DMBELS)”, Int.J.of Computer Applications, vol.66, no.19, pp. 31-36, 2013.

[19] M. AlAjmi, S. Khan, and A. Zamani, “Using Instructive Data Mining Methods to Revise the Impact of Virtual Classroom in E-Learning”, Int. J. of Advanced Science and Technology, vol. 45, pp. 125-134, 2012.

[20] Hung, J., Rice, C., and Saba, A., “ An Educational Data Mining Model for Online Teaching and Learning”. J. of Educational Technology Development and Exchange, vol.5, no.2, pp.77-94, 2012.

[21] T. Chellatamilan and R. Suresh, “An e- Learning Recommendation System using Association Rule Mining Technique”, European J.of Scientific Research, vol.64, no.2, pp.330- 339, 2011.

[22] Mamcenko, J., Sileikiene, I., Lieponiene., “Analysis of E-Exam data using data mining techniques”. available at: http://isd.ktu.lt/it2011/material/Proceedings/6_ITTL_3.pdf

[23] Anitha, A & Krishnan,N., “A Dynamic Web Mining Framework for E-Learning Recommendations using Rough Sets and Association Rule Mining”, International Journal of Computer Applications, vol.12, no.11, pp. 36-41, 2011.

[24] Carmona, J., et.al., “Evolutionary algorithms for subgroup discovery applied to e-learning data”,. available at: http://sci2s.ugr.es/keel/workshops/docs/workshop1/SD_e-learning_ESw.pdf , 2010

[25] Dominguez, A., Yacef, K., and Curran, J. , “data mining for Individualised Hints in eLearning”, 2010.

Fundamental Study to New Evaluation Method Based on Physical and Psychological Load in Care

Hiroaki Inoue
Tokyo University of Science, Suwa
Research course of Engineering/Management
Chino-city, Japan

Shunji Shimizu, Hirotaka Ishihara and Yuuki Nakata
Tokyo University of Science, Suwa
Department of Computer Engineering
Chino-city, Japan

Hiroyuki Nara
Hokkaido University
Graduate School of Information Science and Technology
Sapporo-city, Japan

Takeshi Tsuruga
Hokkaido Institute of Technology
Department of Clinical and Rehabilitation Engineering
Sapporo-city, Japan

Fumikazu Miwakeichi
The Institute of Statistical Mathematics
Spatial and Time Series Modeling Group
Tachikawa-city, Japan

Nobuhide Hirai
Tokyo Medical and Dental University
Health Service Center
Bunkyou-ku, Japan

Senichiro Kikuchi, and Satoshi Kato
Jichi Medical University
Department of Psychiatry
Shimotsuke-city, Japan

Eiju Watanabe
Jichi Medical University
Department of Neurosurgery
Shimotsuke-city, Japan

***Abstract*—In Japan and developed countries, it has become aged society, and wide variety welfare device or system have been developed. But these evaluation methods of welfare device or system are limited only stability, intensity and partial operability. Because of, it is not clear to determine the standard to evaluation for welfare device or system of usefulness. Therefore, we will attempt to establish the standard for evaluation about usefulness for objectively and quantitatively for including non-verbal cognition. We examine the relationship between human movements and brain activity, and consider the evaluation method of welfare devices and systems to measure the load and fatigue which were felt by human. In this paper, we measure the load for sitting and standing movement using NISR. We tried to make sure for the possibility of the quantitatively estimation for physical or psychological load or fatigue by measuring of brain activity using NIRS(Near Infra Red Spectroscopy). As results, when subjects perform the movement task, the statistical significant difference was shown in the specific part of the brain region.**

Keywords—component; Evaluation; Movement; Exercise; NIRS; Care; Welfare Technology; Useful welfare device evaluation; Evaluation method

I. Introduction

As it has been known widely, aging population in Japan and world-wide countries has been increasing. Thereby the number of care worker has been increasing. Care is very hard work. Welfare devices and systems reducing a burden of the care work are required. In this background, welfare systems and device are rapidly developing, and various devices are manufactured based on the increased popularity of welfare device and system. Also, the market of welfare devices and systems is expanding. However, the evaluation method is limited respectively to stability, strength and a part of operability for individual system or device. It means that evaluation methodology for usefulness of them was not established. Therefore, we will attempt to establish a standard to evaluate the usefulness for objectively and quantitatively on the basis of cognition such as physical load, reduction of fatigue and postural stability. Especially, in considering universality, it is necessary to measure human movement in daily life. Movement was not measured by using particular device, but routinely-performed movement in daily life.

In our previous study, we examined the possibility of evaluation by measuring physical load due to activities of daily living with using 3D Motion Analysis System and EMG. Also, we looked into the possibility of quantitative evaluation of tiredness and load on the basis of brain activity using NIRS [1]. We consider that physical and psychological load are linked to cognition including non-verbal cognition. In this paper, our purpose is to obtain an index of the measure of load and fatigue which are felt by human. We discussed experiments which explore the relationship between exercise and human brain activity. In addition, we measured the brain activity of subjects who performed exercise task with imposing additional load. Thus, we tried to measure the load and fatigue felt by human.

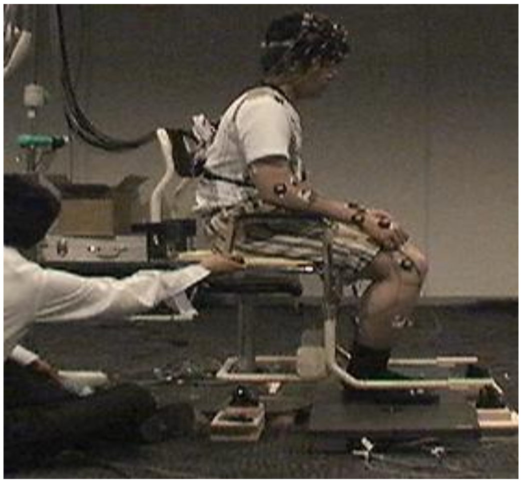


Fig. 1. Experimental View by using NIRS

II. Experimtal Method

A. *Evaluation by using NIRS*

We measured the subject's movement and muscle potential in our previous research [1]. In this paper, we would like to discuss about brain activity when subjects performed the movement tasks which are standing and sitting (Figure 1).

Subjects were six males aged twenty. They were asked to read and sign an informed consent regarding the experiment. Measurement apparatus was NIRS (SHIMADZU CO. Ltd products-FOIRE3000 [4]). Measurement region was at right and left prefrontal cortex.

1) Measuring brain activity during transfer with standing position (task1)

At this measurement, the subjects used welfare device to perform transferring in a standing position. In this measurement, subject sat on seating face of welfare device appeared on the top of chair after raising hip until kneeling position. Also, subject performed inverse transferring from seating face to chair. Time design was rest (5 seconds), task (10 seconds), and rest (5 seconds). This time design was repeated 30 times. Rest time is to stabilize the brain activity. In the measurement NIRS,

2) Measuring brain activity during transfer with half-crouching position (task2)

At this measurement, the subjects used welfare device to perform transferring in a half-crouch position. In this measurement, the subjects sat on seating face of welfare device appeared on the top of chair after raising hip until kneeling position. Also, the subject performed inverse transfer from seating face to chair. Time design was rest (5 seconds), task (10 seconds) and rest (5 seconds). This time design was repeated 30 times.

In experiments of task1 and task2, the operation of welfare device was performed by an operator other than subject. Before this measuring, subjects adjusted to transferring by use of welfare device.

B. *Measuring the brain activity when subjects perform movement task with additional load.*

In this experiment, we have performed measuring brain activity in the case where no load is applied to subject. In previous experiment, load of subjects were derived from only standing and sitting movement. In this experiment, we added the weight to subjects as additional load. As load, there are some cases which are no load 5kg and 10kg. Subjects wear a backpack containing the weight. And Subjects performed standing and sitting movement. Task design was same with previous experiments.

III. Experimental Results

A. *Evaluation by using NIRS*

As the common result of all subjects, oxy-Hb tended to increase during task and to decrease in resting state. Therefore, it was thought that change of hemoglobin density due to task was measured. Fig.3 and Fig.4 show trend of the channel in which significant different was shown. Analysis was performed via one-sample t-test [5,6,7,8,9] by a method similar to previous researches [5,6,7,8,9]. In this analysis, it was necessary to remove other than change of blood flow due to fatigue. So, our method was mainly focused on resting state to compare with the 1st trial and other trials of brain activity.

In task1, 1 and 2, each of sample data for analysis was 4 seconds after the task (Fig. 2).In the t-test of the same task, we performed t-test with first time trial and other trial which was from second times to thirty times, and examined relationship the number of trials and significant differences.

In task 1, significant different could be found from the about 10th trials. Fig. 5 show the region confirmed significant difference. In task 2, significant different could be found from the about 10th trials too. Fig. 6 shows region confirmed significant difference.

At first, we performed t-test using 4 seconds during first trial and 4 seconds during other trials, which were from second to fifteenth in same position.

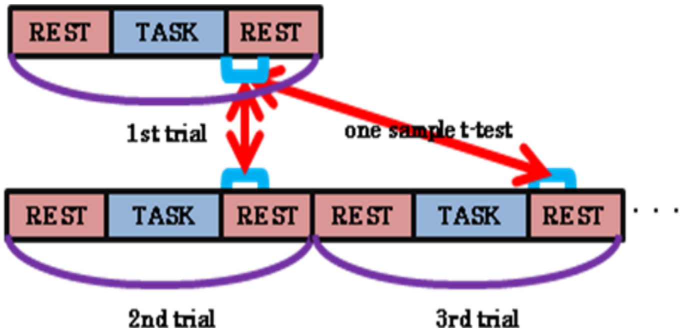


Fig. 2. T-test of sample data in task1 and 2

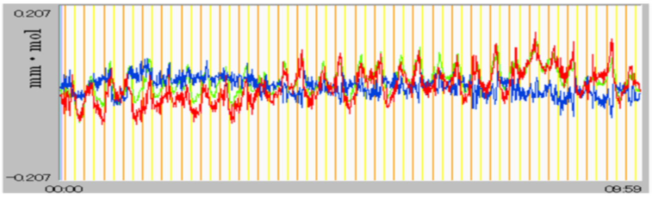


Fig. 3. Measuring result of task1

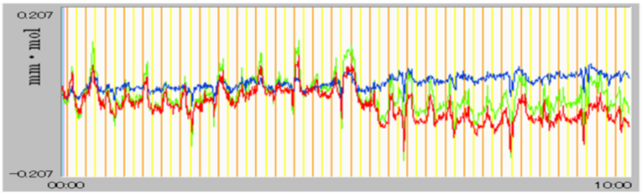


Fig. 4. Measuring result of task2

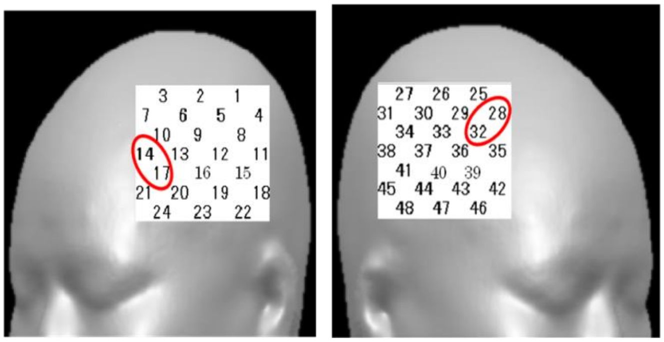


Fig. 5. Signififant difference of task1

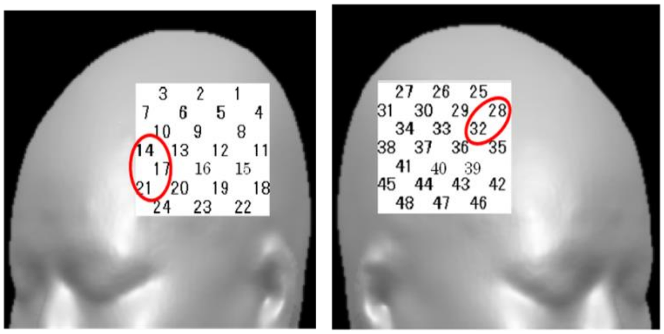


Fig. 6. Significant difference of task2

B. *Results of Brain activity measurements when the load was applied to the subjects*

Analysis method was one-sample t-test of brain activity data as with above analysis. Fig.7 shows the result of one sample t-test between first trial of rest data and another trial of rest data. This analysis method was same with Fig.2. Fig.8 and Fig.9 shows the result of one sample t-test with brain activity data of different movement in the same number of trials. Red circle is the brain region that has seen a statistically significant difference in 5 of 6 subjects. Yellow circle show the brain region that has seen a statistically significant difference in 4 subjects. Green circle show the brain region that has seen a statistically significant difference in 3 subjects.

In the both t-test, there were significant difference on the prefrontal cortex. These results were similar to above experiments. However, significant differences were found randomly regardless of the number of trials. As the cause of these results, there is a possibility that t-test could not remove the changes in scalp of blood flow.

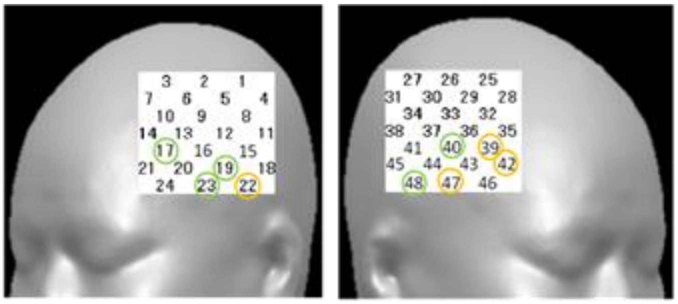


Fig. 7. Result of one sample t-test between first tial and another trials when subjects had no additional load

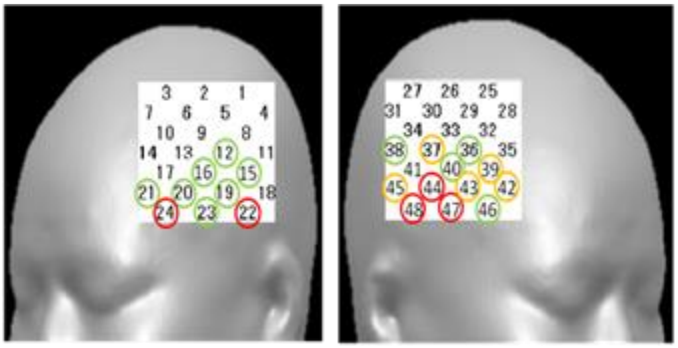


Fig. 8. Result of one sample t-test between first tial and another trials when subjects had 5kg load.

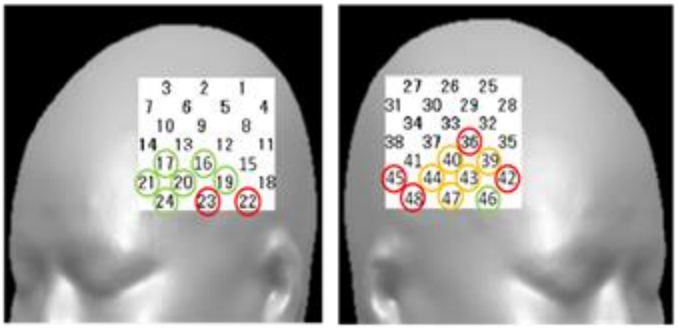


Fig. 9. Result of one sample t-test between first tial and another trials when subjects had 10kg load

IV. DISCUSSION

A. *Evaluation by using NIRS*

In this experiment, we tried to measure quantitatively the physical and psychological strain on the basis of brain activity. Also, we think that brain activity disclose human cognitive including non-verbal. As a result, it was shown that there were differences at brain activity due to number of trials and postural.

In this time, analysis was performed via one-sample t-test using sample of brain activity in resting state during task or after task. Hence, analysis method was to remove disturbance such as body motion and angular variation of neck to the extent possible although there was the possibility to measure skin blood flow. Therefore, it was thought that strain due to tasks was quantitatively measured by being recognized significant differences. Also, in previous research, it was reported to decrease in activity in the brain around #10, 11 [10], as the result of measuring brain activity during Advanced Trial Making Test using PET [11]. Therefore, this result came out in support of previous research in no small part.

B. Results of Brain activity measurements when the load was applied to the subjects

As a step to make the evaluation method, we performed additional experiment, which is imposed to subjects the load other than using the welfare devices. Significant difference was observed in the brain region similar to previous experiments. And, each time the load increases, brain regions found statistically significant differences became widespread. And, the frequency of showing the statistical significant difference became the higher. We think that there are the possibility of happen this results by the additional load.

Of course, it is necessary to increase number of subject at the present stage. In addition, there are problems associated with experiment, number of subject, method and measured region. However, in terms of being recognized significant differences at brain activity due to movement, it was thought to show useful result in evaluating quantitatively daily movements.

V. Conclusion

In this paper, we tried to measure physical and psychological load with measuring brain activity. And there were significant differences due to number of trials. In this experiment, analysis method was to remove disturbance such as body motion and angular variation of neck to the extent possible by using the measurement result in resting state as sample. Therefore, it was thought to show the useful result in evaluating quantitatively load due to movement task by being recognized difference in brain activity caused by number of trials.

Main purpose in this study is to evaluate physical load and fatigue quantitatively. So, we tried to evaluate change of muscle load due to difference of motion by simultaneous measuring with 3D motion analysis System and EMG quantitatively.

However, evaluation of psychological load is necessary, too. In terms of using welfare device, prolonged use must be taken into account. In this case, it is important to consider not only physical load but also psychological load due to prolonged use from standpoint of developing welfare device and keeping up surviving bodily function.

Also, in previous research, separation between physical and psychological load has been performed. But, our view is that there is correlation with physical and psychological load. So, we tried to measure psychological load including physical one based on brain activity and quantitatively evaluate both load.

VI. Future Work

For the future, our purpose is to establish method of discussing useful of welfare device by evaluating load involved in other daily movements with increasing number of subjects. Especially, the investigation to relationship between load of the movement and the brain activity is important in the establishment of the evaluation method to usefulness of welfare devices and systems.

Acknowledgment

This research was supported as follow. There dimensional motion analysis and commodious room were provided from nac Image Technology Inc. EMG and analysis software were provided from KISSEI COMTEC. NIRS was provided from SHIMADZU COPPORATION. Developing welfare device was provided from IDEA SYSTEM CO., LTD. We are deeply grateful to them.

This study contributes to become the basis for one of theme of s-innovation program in Japan Science and Technology Agency which was named "Development Fatigue-reduction Technology for Social Contribution of Aged Person and Establishment System for Evaluation.

References

[1] H.Inoue, S.Shimizu, N.Tkahashi, H.Nara, T.Tsuruga, F.Miwakeichi, N.Hirai, S.Kikuchi, E. Watanabe, S.Kato," Fundamental Study to New Evaluation Method Based on Physical and Psychological Load in Care,", IARIA cognitive2012, Nice, France, pp101-106.

[2] Y. Shinoda, "Considertion of feature extraction based on center of gravity for Nihon Buyo dancer using motion capture system," SICE Annual Confference 2011, Tokyo Japan, pp. 1874-1878.

[3] Y. Yamaguchi, A. Ishikawa, and Y.Ito, "Development of Biosignal Integration Analysis System for Human Brain Function and Behavior," Organization for Human Brain Mapping 2012, China.

[4] H.Inoue, S. Shimizu, N. Takahashi, H.Nara, and T. Tsuruga "Fundamental Study for Evaluation of the Effect due to Exercise Load" , Assistive Technology, Bio Medical Engineering and Life Support 2011, Japan.

[5] E. Watanabe, Y. Yamashita, Y. Ito, and H. Koizumi, "Non-invasive functional mapping with multi-channel near infra-red spectroscopic topography in humans," Heurosci Lett 1996, Feb 16, 205(1), 41-4.

[6] N. Takahashi, S. Shimizu, Y. Hirata, H. Nara, F. Miwakeichi, N. Hirai, S. Kikuchi, E. Watanabe, and S. Kato, "Fundamental Study for a New Assistive System during Car Driving," Proc. of International Conference on Robotics and Biomimetics, DVD-ROM, 2010, Tenjin, China.

[7] N. Takahashi, S. Shimizu, Y. Hirata, H. Nara, H. Inoue, N. Hirai, S. Kikuchi, E. Watanabe, and S. Kato, "Basic study of Analysis of Human Brain Activities during Car Driving," the 14th International Conferrence on Human-Computer Interaction, 2011, USA.

[8] S. Shimizu, N. Takahashi, H. Nara, H. Inoue, and Y. Hirata, "Fundamental Study for Human Brain Activity Based on the Spatial Cognitive Task," the 2011 Internatinal Conference on Brain Informatics-BI 2011, China.

[9] S. Shimizu, N. Takahashi, H. Nara, H. Inoue, and Y. Hirata, "Basic Study for Human Brain Activity Based on the Spatial Cognitive Task," The Third International Conference on Advanced Cognitive Techonologies and Applications, 2011, Italy.

[10] S. Shimizu, N. Takahashi, H. Inoue, H. Nara, F. Miwakeichi, N. Hirai, S. Kikuchi, E. Watanabe, and S. Kato, "Basic Study for a New Assitive

System Based on Brain Activity associated with Spatial Perception Task during Car driving," Proc. International Conferrence on Robotics and Biomimetics, 2011, Thailand.

[11] Y. Watanabe,"Molecular/neural mechanisms of fatigue, and the way to overcome fatigue," Folia Pharmacological Japonica, vol. 129, pp. 94-98, 2007.

[12] H. Kuratsune, K. Yamaguti, G. Lindh, B. Evengard, G. Hagberg, K.Matsumura, M. Iwase, H. Onoe, M. Takahashi, T. Machii, Y.Kanakura, T. Kitani, B. Langstrom, and Y. Watanage,"Brain Regions Involved in Fatigue Sensation: Reduced Acetylcarnitine Uptake in to the Brain," Neuroimage, vol. 17, pp. 1256-1265, November 2001.

[13] K. Maruta,"The influence of Seat Angle on Forward Trunk Inclination During Sit-to-Stand," Jounal of Japanese Physical Therapy Association, vol. 31, No.1, pp. 21-28, 2004.

Improved Security of Audit Trail Logs in Multi-Tenant Cloud Using ABE Schemes

Bhanu Prakash Gopularam
Cisco Systems India Pvt. Ltd
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India

Nalini N
Department of Computer Science and Engineering
Nitte Meenakshi Institute of Technology
Bangalore, India

***Abstract*—Cloud computing is delivery of services rather than a product and among different cloud deployment models, the public cloud provides improved scalability and cost reduction when compared to others. Security and privacy of data is one of the key factors in transitioning to cloud. Typically the cloud providers have a demilitarized zone protecting the data center along with a reverse proxy setup. The reverse proxy gateway acts as initial access point and provides additional capabilities like load balancing, caching, security monitoring capturing events, syslogs related to hosts residing in the cloud. The audit-trail logs captured by reverse proxy server comprise important information related to all the tenants. While the PKI infrastructure works in cloud scenario it becomes cumbersome from manageability point of view and they lack flexibility in providing controlled access to data. In this paper we evaluate risks associated with security and privacy of audit logs produced by reverse proxy server. We provide a two-phase approach for sharing the audit-logs with users allowing fine-grained access. In this paper we evaluate certain Identity-Based and Attribute-Based Encryption schemes and provide detailed analysis on performance.**

***Keywords*—multi-tenancy; audit-trail log; attribute-based encryption; reverse proxy security**

I. Introduction

Cloud computing as defined by NIST is a model for enabling convenient, on-demand network access to a shared pool of configurable resources that can be rapidly provisioned and releases with minimal management effort or interaction. While the private cloud gives organizations greater control over the infrastructure it may not be cost effective for small and medium businesses [1]. Cloud services are offered in different service models and three well known models are Infrastructure as a Service – Cloud user has greater control on infrastructure Vmware, Openstack, Azure offer such services. Platform as a Service – developer centric services Heroku, Google AppEngine are few providers, and Software as a Service – services include data analytics, online meetings such as Cisco WebEx, Gmail. Cost reduction and increased efficiency are primary motivations towards a public cloud and nevertheless security and privacy objectives play vital role for decisions about outsourcing IT services [2]. The data collected by network devices such as firewalls, reverse proxy servers, hypervisor are very vital for monitoring health of cloud as well as for security forensics [3].

The internet facing reverse proxy gateway provides protection from issues like intrusion detection, denial of service attacks etc. Data collected by reverse proxy includes system logs, alarms and it can capture HTTP/REST requests, remote-service calls pertaining to tenants if it is configured as SSL termination end-point. The traditional way of log sharing suffers from few problems like:

1) Existing PKI based techniques for preserving the audit logs largely relay on certificates for exchanging the data. Considering cloud storage as untrusted, managing centralized repository for certificates of multitude of tenants necessities frequent synchronization with key servers and the process is error-prone due to large number of interactions with PKG server.

2) The traditional PKI infrastructure either reveals all the data or restricts and does not provide easy way to allow fine-grained access to data considering organizational policy information.

II. Key Contributions

In this paper we outline challenges associated with audit-log preservation in cloud with reverse proxy architecture. We experiment with advances in attribute-based encryption schemes to overcome privacy and security problem of audit trail logs. We experiment with Identity-based encryption techniques proposed in [7] referred as *BB* scheme here by and *committed blind anonymous* identity-based encryption as proposed in [8] referred as *CKRS* scheme here by.

Ciphertext Policy Attribute Based Encryption techniques proposed in [9] referred as *BSW* scheme here by and *efficient and secure realization* of CP-ABE scheme proposed in [10] referred as *Waters* scheme here by.

In literature the above schemes are also considered as key milestones in the identity-based cryptography. We provide a work flow for secure distribution of audit-trail logs captured by reverse proxy server among multiple tenants. We evaluate the performance of operations like setup, key-generation, encryption, decryption under various configurations. Few applications of proposed scheme include secure reverse proxy implementation without overhead of certificate management, enabling on-demand third party auditing or inspection, and secure sharing of logs with interested parties with fine-grained access control.

III. Preliminaries

A. Identity and Attribute Based encryption:

Shamir first proposed concept of Identity-based public key cryptography in mid 1980 and in 2001 the first practical and secure IBE scheme [12] was presented by Boneh and Franklin. Sahai and Waters [14] first introduced concept of Attribute-Based encryption in 2006, the user attributes were used to encrypt and decrypt data. In the same year Bethencourt et al. [9] presented first construction of Ciphertext-Policy Attribute-Based encryption. Using CP-ABE it is possible to embed role based access control policies into the ciphertext. Later attribute based encryption was extended to distributed identities and hierarchical attribute based encryption schemes. ABE systems now support many crucial functionality [15] required by security infrastructure.

B. Reverse Proxy Gateway

A reverse proxy is server side software typically acts as entry point for HTTP requests. Typically reverse proxy resides in DMZ facing the internet. The HTTP request is scrutinized first and requested content is served if it is already in cache or statically referred. This setup is compelling for cloud service providers with multi-tenancy architecture and having a single entry point with capability to route requests provide lots of benefits for CSPs. Other important usecases include B2B transactions, supply chain integration. Some of the key functionality provided by reverse proxy gateway architecture is described here. One can refer reverse proxy websites like Nginx [4], SkyHigh software architecture to know more.

a) Security: Reverse proxy can provide single point of communication. It can decrypt the HTTPS based request and communicate with back end servers in HTTP mode. Provides many advantages for cloud users like ease of configuration of SSL/TLS, saves CPU intensive security operations using specialized hardware.

b) Centralized Logging and Auditing: As all HTTP requests are routed through reverse proxy server, it captures all the important events related to hosts residing in the cloud.

c) Load balancing: RP can route the incoming HTTP requests among the available servers using strategies like round robin, sticky session in case of stateful sessions etc.

d) Caching and serving static content: For storage based cloud applications viz., youtube, vimeo the server responsiveness can be improved by hosting static content and using RP for routing.

C. Audit Trail Log Structure

Reverse proxy can be configured to generate logs like file, stderr or syslogs. For example in Nginx server, the log format is specified using log_format directive

```
http {
 log_format compression
'$remote_addr - $remote_user [$time_local] '
"$request" $status $body_bytes_sent '
'"$http_referer" "$http_user_agent";
}
```

TABLE I. AUDIT-TRAIL LOGS GENERATED BY REVERSE PROXY GATEWAY

Attribute Name	**User Login Activity**	**Resource Access Activity**
Time	14:14:19.566	12:13:26.080
UserID	Supervisor801	admin
EventType	User Logging	User Access
EventStatus	Failure	Success
ClientAddress	https:64.103.237.53 :tcp:54665	64.103.237.53
ResourceAccessed	AppAdmin	Channel Provider/2
CompulsoryEvent	Yes	No
ComponentID	Administration	Configuration API
AuditCategory	Authentication attempt failed	channelProvider/2 2 modified
AppId	10023	1055
ClusterId	1	1
NodeId	uccx-93-55	uccx-93-55

IV. CHALLENGES IN PRESERVING AUDIT-TRAIL LOGS IN MULTI-TENANCY CLOUD

Besides many potential benefits the public cloud the data security is complex due to following challenges

- *Shared Multi-tenant Environment* – Public cloud achieves multi-tenancy by logical separation at multiple layers of software stack. The attacker can pose as a consumer and exploit vulnerabilities from cloud environment.
- *Loss of Control* – While the cloud users may perceive the services as traditional service model, transitioning of control to cloud provider amplifies the risks associated.
- *System complexity* – Complexity largely depends on infrastructure used and often the cloud providers use methods that are proprietary in nature. Typically complexity relates inversely to security, the complexity leads to increased risk for vulnerabilities.
- *Audit trail logs* – the cloud computing environment poses new challenges from audit and monitoring perspective. Full audit trail within the cloud is still an open problem and poses lots of challenges as seldom organization security policy challenges does meet the cloud provider practices.

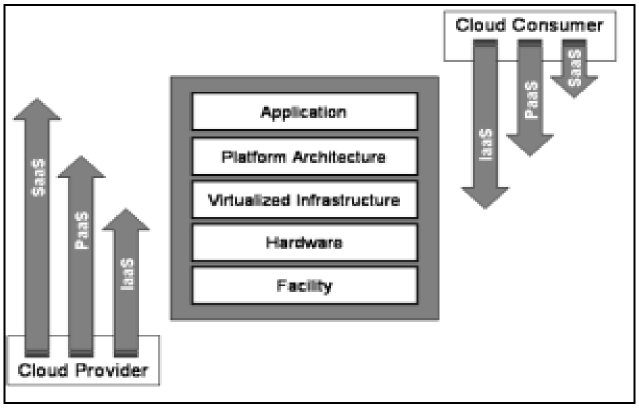


Fig. 1. Cloud service models and differences in scope and control

V. METHODOLOGY

Consider public cloud provider having multiple tenants and protected with reverse proxy server which captures audit-trail records of incoming traffic. We consider role of reverse proxy server extended as SSL termination end-point so that it can intercept all HTTP/SSL traffic. The cloud provider has a Network Admin who has access to entire logs and cloud tenants with users having roles like level-1, level-2, level-3 etc. While level-1 users are in the bottom of organizational hierarchy and they are monitored by level-2 and so on and so forth.

A. Privacy and Security of Audit logs - Objectives

We divide the problem into two sub-domains – 1. Cloud Network Admin has access control on entire logs and can do operations like key-generation, encryption, decryption, 2. Tenant users like Network Admin can access all tenant specific logs and users of Level-1, Level-2 etc. has controlled access to data. Users at higher level can oversee data pertaining to lower level that they are administering. It implies that user's access to audit log contents is controlled using *role-based access control* policies.

TABLE II. CLOUD USERS ACCESS TO CONTENTS OF AUDIT LOGS CATEGORIZED INTO TYPE-1, TYPE-2 SECURITY

Participant Role	Accessible content in audit-trail log	Category
Level-1 [Tenant]	Time, UserID, EventType	Type-1
Level-2[Tenant]	Time, UserID, EventType, EventStatus, ClientAddress, ResourceAccessed	
Network Admin [Tenant]	Time, UserID, EventType, EventStatus, ClientAddress, ResourceAccessed, CompulsoryEvent, ComponentID	
Cloud Network Admin [Cloud Provider]	Time, UserID, EventType, EventStatus, ClientAddress, ResourceAccessed, CompulsoryEvent, ComponentID, AuditCategory, AppId, ClusterId, NodeId	Type-2

B. Design

For audit-trail log security we choose 2-phase protection. The unique challenge here is that the security mechanisms should ensure that cloud providers has complete control on the data and has ability to share with tenants and while restricting access according to organizational hierarchy. We solve this problem using blend of identity and attribute-based encryption schemes. The problem is solved in 2-phase approach

a) Type-I Data Security

Type-I data protection involves security mechanism like Identity-based encryption [7] [8]. The Cloud Network Admin has access to all the data but individual tenants should have access to their data only. We use identity-based encryption scheme for access control. Each encrypted log entry is associated with public identifiers or tags like *TenantId* and user keys are associated with access policy. Although entire logs are kept in shared location in cloud, the individual tenants can access only their data. The reason for choosing identity-based encryption scheme is that it is possible to share data without requiring exchange of certificates. We evaluated two identity-based encryption schemes [7] [8] for performance with large datasets.

b) Type-II Data Security

Type-II data security involves allowing fine-grained access control on data to tenant users. The user can decrypt data only if the attributes in secret key satisfies the access structure of encrypted data. For example Level-1 user can see only her own activity while the Level-2 can see activity of all his employees and soon. We use ciphertext-policy attribute based encryption schemes [9][10] with ciphertext having policy information of participants and user keys having descriptive attributes about participant. The reason for choosing CP-ABE scheme here is that it is perfectly suited for environment where user privileges (*role-based access control*) determine the access to data and it allows fine-grained access control on the data. We experiment with *BSW* [9] and *Waters* scheme [10] for performance.

Depending on the log sharing mechanism two possible approaches exist.

- Cloud provider use Type-1 security mechanism for logs encryption and Cloud tenants access their data and decrypt and re-encrypt using Type-2 security mechanism
- Cloud provider applies Type-2 security mechanism which internally uses policy tree for log encryption and then re-encrypt using Type-1 security mechanism. The tenants access the data by using Type-1 secret keys and then use Type-2 secret keys to decode the data.

In this paper we provide experimental results of Type-1 and Type-2 security mechanisms separately and results are applicable in both the cases outlined.

c) Setup and Key Generation

- Type-1: The algorithm initialization depends on bilinear pairing and elliptic curve used. The master secret *MK* and public key *PK* are generated using system parameters *P*.

- Type-2: This can be done by cloud provider or tenant itself depending on use case. The CP-ABE *Setup(k)* is run with security parameter and it results in public parameters (*PK*) and master key (*MK*). The CP-ABE KeyGen(*MK*, *PK*, T) with possible tenant-id values which outputs decryption keys associated with attributes.

d) Encryption and Decryption:

- Type-1: Each log entry, the data <ApplicationId, ClusterId, NodeId> is encrypted using a symmetric key algorithm and using individual tenant-id t_i as public key the server computes ciphertext c_i of the data and *P* as public parameters. Here *P* may be equal to t_i if cloud provider wishes to annotate with tenant-id only. Data is decrypted using (*PK*, *sk*, CT)

- Type-2: For each log record, the data pertaining to tenants, the Enc(*Record*, T, *PK*) where T relates to access structure T for public parameters *PK*. The CP-ABE Dec(CT, *SK*, *PK*) is run using user secret keys *SK* and public parameters *PK*.

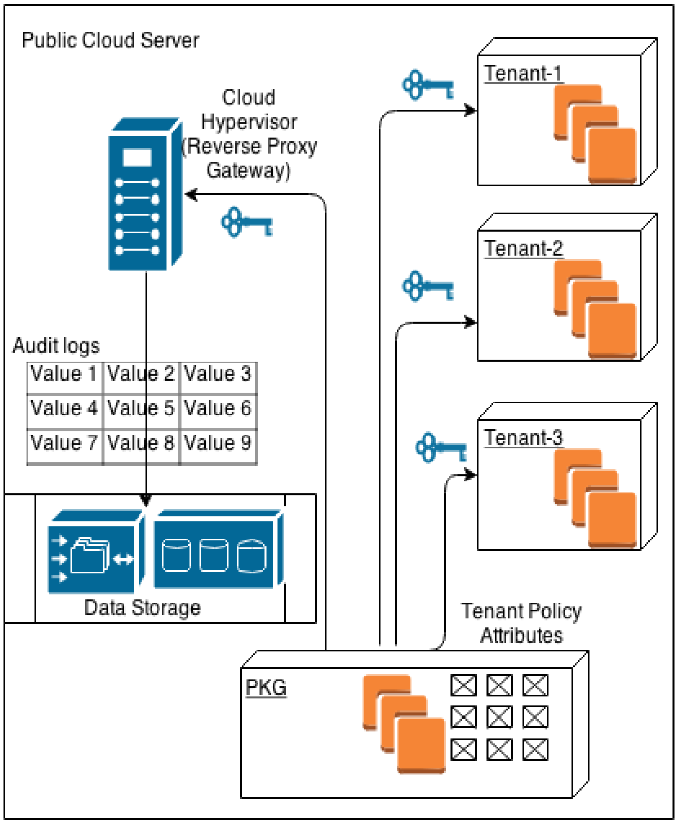


Fig. 2. Multi-Tenant cloud with audit trail mechanism secured using combination of ABE

VI. Experiments and Evaluation

We use a hypothetical example of public cloud provider hosting 3 tenants.

A. *System Details-We have used CHARM crypto-library[5] [6] v0.43 for prototyping. At a very high-level the library provides a protocol engine for many cryptographic operations and an adapter architecture which bridges gaps necessary for building a complete crypto system. In addition we used other open source libraries including OpenSSL 1.0.1, GMP 6.0.0a and Pairing-Based Cryptography library version 0.5.14 of Stanford. The experiments were carried on X86 based platform using Ubuntu 12.04.4 LTS (precise) 32-bit server with 8 GB RAM and Intel Core i5-3470 CPU with 3.2 GHz 4 core processor.*

B. *Test Data - The sample audit-trail logs used in experiments is sampled from a reverse proxy server. The dataset is split into chunks of approximately 20000 records carefully having activity of cloud tenants with possible operations. We analyze performance of cryptographic schemes with these chunks.*

C. *Data Security – Results - We have used elliptic curve with bilinear maps (or pairings) like 512 bit symmetric curve. We used Type-A curve such as $y^2 = x^3 + x$ to compute the pairings. The secret key is communicated to interested parties using a secure channel like TLS/SSL*

TABLE III. Setup Time for Type-1 Security (IBE Schemes)

Operation	**Scheme**	**Time (milliseconds)**
Setup	Ibe-bb[a]	15.624
	Ibe-ckrs[a]	52.361

a. For pairings symmetric curve with 512 bit is used

TABLE IV. Key Generation Time for Type-1 Security

Operation	**Scheme**	**Time (milliseconds)**
Key Generation	Ibe-bb	3.137
	Ibe-ckrs	22.689

While implementing Type-1 security, the *CKRS* scheme took time prohibitively large time than *BB* scheme for initial setup (master key and public key generation) and secret key generation.

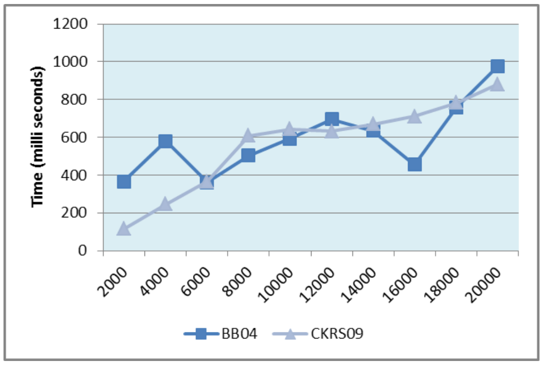


Fig. 3. Encryption using Type-1 (IBE schemes)

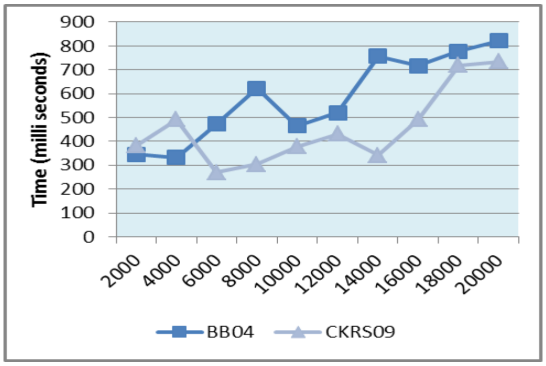


Fig. 4. Decryption using Type-1 (IBE schemes)

The encryption and decryption of data using *CKRS* scheme was more performant (5-10%) then *BB* scheme with large datasets. The encryption involves generating a random symmetric key using pairing and encrypting the data using symmetric crypto system such as AES in CBC mode with 16 byte block size. Then the symmetric key is encrypted using IBE algorithm.

D. Type-2 Data Security – Results

The initial setup time for *BSW* scheme was approximately twice the *WATERS* scheme initialization. And secret key generation time for level-1 and level-2 users of cloud tenant with *BSW* and *WATERS* scheme was roughly same.

TABLE V. SETUP TIME FOR TYPE-2 SECURITY (CPABE SCHEMES)

Operation	**Scheme**	**Time (milliseconds)**
Setup	cpabe-bsw[b]	38.305
	cpabe-waters[b]	21.2

b. For pairings symmetric curve with 512 bit is used

TABLE VI. KEY GENERATION TIME FOR TYPE-2 SECURITY

Operation	**Scheme**	**Level-2 key**	**Level-1 key**
Key generation	cpabe-bsw	23.339	23.467
	cpabe-waters	24.569	24.404

The *BSW* scheme took comparatively more time for encryption and decryption with large datasets than *WATERS* scheme and *WATERS* scheme performance was quite stable.

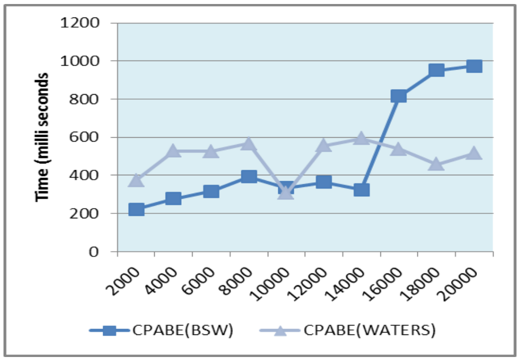


Fig. 5. Encryption using Type-2

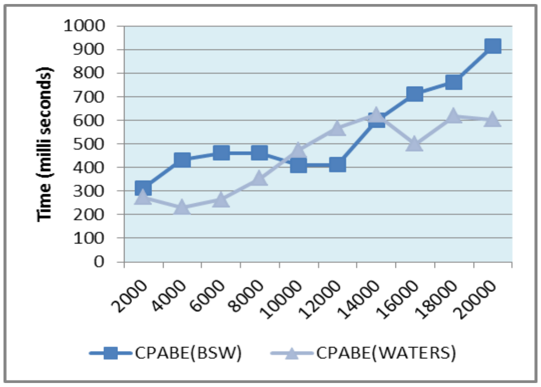


Fig. 6. Decryption by Level-2 tenant user using Type-2

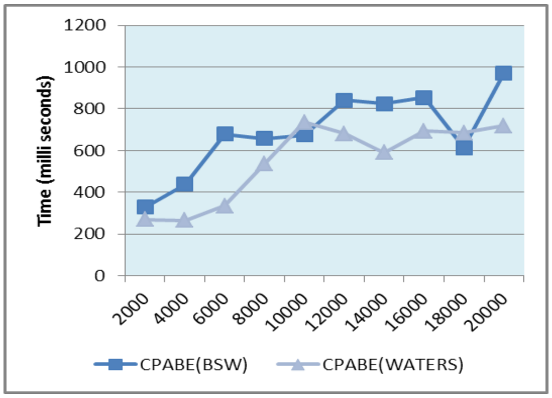


Fig. 7. Decryption by Level-1 tenant user using Type-2

The proposed scheme provides security of sensitive data and provides fine-grained access control to the data and following are some limitations of identity based systems.

1) A unique characteristic of identity based systems that differentiates from existing PKI schemes is that the encryption is possible without any need for communicating with server during validity period of the public parameters. This reduces network communication significantly but can lead to problems in case of lost key or key revocation. Given the fact that identity based systems require lesser validation with key servers, the IBE private keys should not be created using timestamp of longer duration as this could worsen the problem of key compromise.

2) Using identity system based on bilinear pairings devised on family of supersingular elliptic curves over finite fields (also called as type-1 curves) for practical applications using Advanced Encryption Standard keys it is sufficient to use 128-bit levels and higher such as 192 bits or 256 bits.

VII. CONCLUSION AND FUTURE WORKS

Audit log preservation in cloud is a challenging problem considering the dynamicity of cloud. The current mechanisms to share the logs securely involve large overhead in terms of certificate management and do not offer flexibility to share data. The combination of different identity-based encryption techniques discussed in this paper provide a simpler mechanism for log-sharing to intended receivers. In future we plan to extend the research to implementing oblivious search on encrypted audit logs along with computation on data like analytics with monotonic and non-monotonic access structures and along with predicate encryption.

ACKNOWLEDGMENT

We would like to thank Sashank Dara and *crypto.stackexchange* community for insightful discussions and suggestions. We would also like to thank Satheesh Kumar and Sridhar Gaddipati, CBABU, Cisco Systems for supporting us while carrying this work.

REFERENCES

[1] Michael Armbrust et al, "Above the Clouds: A Berkeley View of Cloud Computing", Technical Report No. UCB/EECS-2009-28, February 10, 2009

[2] Siani Pearson and Azzedine Benameur, "Privacy, Security and Trust Issues Arising from Cloud Computing", 2nd IEEE International Conference on Cloud Computing Technology and Science, HP Labs, pp. 693-702

[3] Binti Abdul Aziz, N, Binti Meor Yusoff, N.D, Binti Abu Talib, "Log Visualization of Intrusion and Prevention Reverse Proxy Server against Web Attacks", Informatics and Creative Multimedia (ICICM), International Conference, 2013, pp. 325-329

[4] Wei Yuan, Hailong Sun, Xu Wang, Xudong Liu, "Towards Efficient Deployment of Cloud Applications through Dynamic Reverse Proxy Optimization", High Performance Computing and Communications, IEEE, 2013, pp. 651 - 658

[5] Yannis Rouselakis, Brent Waters, "Practical constructions and new proof methods for large universe attribute-based encryption", ACM SIGSAC conference on Computer & communications security, 2013, pp. 463-474

[6] Joseph A Akinyele, Christina Garman, Ian Miers, Matthew W Pagano, Michael Rushanan, Matthew Green, Aviel D Rubin, "Charm: A framework for rapidly prototyping cryptosystems", Journal of Cryptographic Engineering, Springer-Verlag, 2013, pp. 111-128

[7] Dan Boneh , Xavier Boyen, "Efficient Selective-ID Secure Identity Based Encryption Without Random Oracles", Proceedings of Eurocrypt 2004, volume 3027 of LNCS, 2004, pp. 223-238

[8] Jan Camenisch , Markulf Kohlweiss , Alfredo Rial , Caroline Sheedy, "Blind and Anonymous Identity-Based Encryption and Authorised Private Searches on Public Key Encrypted Data", PKC 2009

[9] John Bethencourt, Amit Sahai, and Brent Waters, "Ciphertext-policy attribute-based encryption", 28th IEEE Symposium on Security and Privacy, Oakland, May 2006 , pp. 321-334

[10] Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in Public Key Cryptography - PKC 2011 , Vol. 6572, 2011, pp. 53–70.

[11] B. R. Waters, D. Balfanz, G. Durfee, and D. K. Smetters, "Building an encrypted and searchable audit log", 11th Annual Network and Distributed System Security Symposium 2004

[12] Boneh, D., Franklin, M, "Identity-Based Encryption from the Weil Pairing", Kilian, J. (ed.) CRYPTO 2001. Springer LNCS, vol. 2139, pp. 213–229

[13] V. Goyal, O. Pandey, A. Sahai and B. Waters, "Attribute Based Encryption for Fine-Grained Access Conrol of Encrypted Data", ACM conference on Computer and Communications Security, 2006

[14] Sahai and B. Waters, "Fuzzy Identity Based Encryption", IACR ePrint Archive, Report 2004/086

[15] Schridde, C, Dornemann, T, Juhnke, E, Freisleben, Smith, M."An identity-based security infrastructure for Cloud environments", IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS), 2010 pp. 644-649

A System Supporting Qualitative Research

Emilia Todorova
St Cyril and St Methodius University
of Veliko Turnovo,
Veliko Turnovo, Bulgaria

Dimo Milev
St Cyril and St Methodius University
of Veliko Turnovo,
Veliko Turnovo, Bulgaria

Ivaylo Donchev
St Cyril and St Methodius University
of Veliko Turnovo,
Veliko Turnovo, Bulgaria

***Abstract*—this paper presents a system aimed to be entirely located in the Internet. The database, server side, logic and user interface are accessible regardless the location and equipment of the working team members. The system supports qualitative analysis of large datasets, and the work can be distributed among several team members or teams. A modified mixed method for developing software projects is used, according to the peculiarities of our case, splitting the phase of code writing into two sub-phases each with different number of iterations.**

Keywords—Information systems (IS); Qualitative data analysis (QDA); NoSQL Database; Software Design; Computer Assisted Qualitative Data Analysis Software (CAQDAS)

I. Introduction

The ability to easily gather capacious information on various issues results on one hand results in more accurate and complete data, but on the other needs a relatively easy and convenient way to categorize, evaluate and extract most precise results from the data. Handling massive amounts of data and their processing often dramatically slows down the work. At the same time, attempts to achieve most accurate results require collection of maximum amount of data. Nowadays with the help of information technologies gaining huge amounts of data is an easy task. The hard part is categorization, evaluation and overall extraction of useful information from collected data – in other words, qualitative analysis.

The existence of a software system for organization of qualitative data analysis is of tremendous help to the team1s involved in qualitative research. "Computer assisted data analysis software can be used to make the research process more transparent." [1].

The purpose of this work is to present a cloudy-based system supporting qualitative data analysis. The database, server side, logic and user interface will be accessible regardless to the location and equipment of the working team members.

The system is developed by three groups of graduate students from the Master Course in Information Systems, consulted by lecturers and an expert from an IT company.

II. Defining the Goals

The creation and handling of a system to facilitate qualitative data analysis sets new tasks to developers as well as to trainers in this field. We set three main goals in our work.

The main goal was to develop an on-line system supporting qualitative data analysis, located entirely on the Internet.

The second goal was to explore different methods for developing software projects and choose or propose an appropriate one.

The third goal was to involve in the project Master course students, teachers and IT professionals, and to evaluate the teaching and learning outcomes.

III. Prerequisites

A. Qualitative Research

Some authors emphasize on the research purpose and focus: qualitative researchers are interested in understanding the meaning people have constructed, that is, how people make sense of their world and the experiences they have in the world [2].

Others consider an epistemological point of view: qualitative research is research using methods such as participant observation or case studies which result in a narrative, descriptive account of a setting or practice. [3]

A third group of definitions concentrates on the technology and context of data collection: qualitative research is a situated activity that locates the observer in the world. It consists of a set of interpretive, material practices that makes the world visible. These practices transform the world. They turn the world into a series of representations, including field notes, interviews, conversations, photographs, recordings, and memos to the self. At this level, qualitative research involves an interpretive, naturalistic approach to the world. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or to interpret, phenomena in terms of the meanings people bring to them [4].

Yet there is a simpler and more functional definition suggested by Nkwi, Nyamongo, and Ryan [5]: “Qualitative research involves any research that uses data that do not indicate ordinal values.”

B. Qualitative data

Qualitative data can be arranged into categories that are not numerical. These categories can be physical traits, gender, colors or anything that does not have a number associated to it. Qualitative data is sometimes referred to as categorical data [6].

The project is supported by the Bulgarian Ministry of Education, Project RD-672-08/09.05.2012.

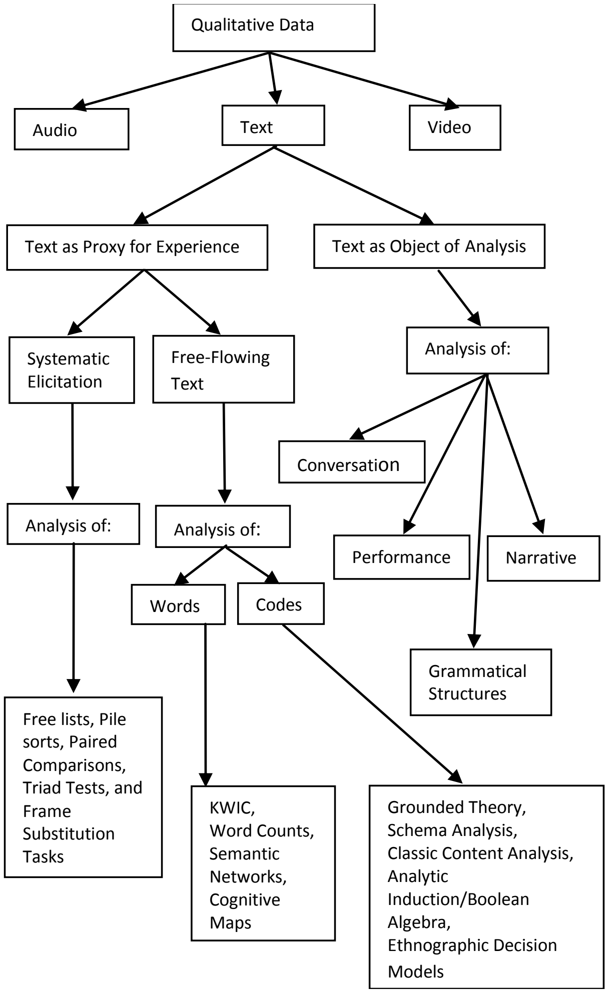

Qualitative Data
Audio
Text
Video
Text as Proxy for Experience
Text as Object of Analysis
Systematic Elicitation
Free-Flowing Text
Analysis of:
Conversation
Analysis of:
Analysis of:
Performance
Narrative
Words
Codes
Grammatical Structures
Free lists, Pile sorts, Paired Comparisons, Triad Tests, and Frame Substitution Tasks
KWIC, Word Counts, Semantic Networks, Cognitive Maps
Grounded Theory, Schema Analysis, Classic Content Analysis, Analytic Induction/Boolean Algebra, Ethnographic Decision Models

Fig. 1. Qualitative data types [7]

Qualitative data is data that approximates and characterizes but basically doesn't measure the attributes of a thing or a phenomenon. It describes data as compared to quantitative data that tends to calculate data. Qualitative has to do with quality and can be subjective. Quantitative has to do with quantity and is measured in numbers.

IV. THE METHOD

The implementation of the software project (information system with integrated database) requires determination of the method of construction. There are two basic styles for developing software projects [8]. One of them is the waterfall method, when the project is partitioned into phases according to the basic activities that are carried out successively. The iterative style sections the project into parts on the base of functional subsets. If we agree that the full software life cycle consists of analysis, design, code writing, testing and evaluation, the first style supposes these phases to be carried out once each, one after another, and the second many times each until the desirable full set of functionalities is reached.

After examining the advantages and disadvantages of existing methods and research experience in this area, we outlined some trends. Most developers especially those involved in object-oriented design do not support the waterfall approach. One of the most important reasons is the difficulty to determine whether the project is developing properly at intermediate stages. It is considered that when using iterative approach the iterations should have fixed time duration (time boxing). When appears that for this period is not possible to finish the planned activities, it is necessary to leave some of the functionalities for the next iteration. This is also a way to prioritize functionalities [8]

For this software project a mixed method is adopted (Fig. 2), The step of writing code is divided into two parts running in parallel: development of the skeleton of the information system and writing modules, implementing specific tools for qualitative data analysis. The special feature here is that the number of embodiments of these two parts may vary prior to the step of testing, such as the expected number of iterations for writing the code of the skeleton of the IS to be less than the number of iterations for the implementation of the functionalities in the qualitative analysis. Naturally it could be expected after a certain iteration skeleton of the information system to be completed, and the inclusion of new tools for qualitative data analysis continues, which means upgrading the functionality of the system.

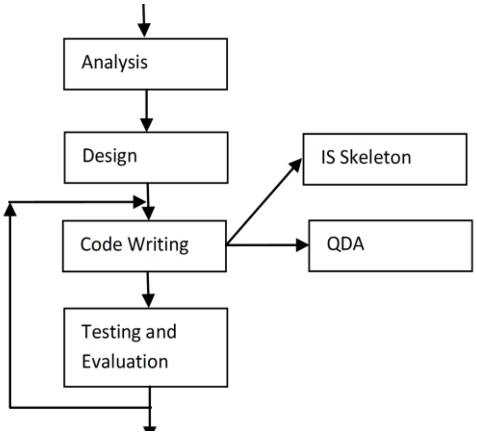


Fig. 2. Method for Developing the Software Project

Determining the iterations of the software project allows separating groups for the development of each phase. There are three groups in the team: the first implements the analysis and design of high level; the second writes code skeleton of the application; the third writes code for the instruments of qualitative data analysis. All groups are involved in the testing and evaluation of the results.

V. IMPLEMENTING THE SYSTEM

A. Defining the Functionalities

The group of analysts examined several CAQDAS packages: Ethnograph, QSRNvivo, Atlas.ti, QDA Miner, Dedoose, DRS (Digital Replay System), NyperRESEARCH, MAXQDA.

The observation and analysis of the methods of qualitative analysis and existing systems helped to define the functionalities of the desirable system.

Functionalities available to users of the system - each user has the following options:

- launch a project for which becomes a manager;
- participate in projects as an analyst, and in this case may assign tags and grades to the paragraphs of the project;
- in case the user has the rights of a project manager he/she can evaluate the work of analysts and edit their grades and labels;

Functionalities provided to Project Manager: each project manager has the following options:

- create a database of paragraphs - one by one or by importing data;
- join or remove members of the team (analysts) to their projects and their assigned rights - analyst, manager;
- create descriptor paragraphs - typical paragraph descriptions that subsequently may be used to group data and analysis results;
- create labels and sub-labels for a project;
- create ratings for the labels;
- delete paragraphs - individually or as a group depending on a specific tag or attribute group (label and grade);
- review the work of each user connected separately and to assess;
- retrieve various reports and graphs based on the processed data;

Features of projects and paragraphs labels:

- database of texts separated into paragraphs;
- tags that analysts assign to paragraphs;
- grades assigned to the labels of a paragraph.

Paragraphs in turn have:

- descriptors - fields describing the paragraph;
- labels and sub-labels;

Labels have grades.

B. The Architecture

The system is built according to the multilayer model architecture. This model is considered to be one of the templates of software architecture. It represents the client-server concept, where the user interface (environment), logic application and work with the database are performed by independent modules. An advantage is that the three-layer model allows to change technology and modify each of the modules without affecting the other layers. The only requirement is to abide by the approved interfaces among the modules. It is admissible the middle (logical) layer to be multi-layer itself, which results in an N-layer model. Fundamental rule in this type of architecture is that the presentation layer (user environment) never communicates directly with the database layer. This makes the model linear.

C. Technologies for Inplementing the Levels

Web technologies are currently the most rapidly developing information technology from the point of view of global connectivity and cloud computing. New technologies in this field appear literally every day and a web programmer needs to stick on the trends. The JavaScript language is well known for Web developers for it is relatively simple but sufficient and powerful. Until recently it was used on1ly for building user interfaces, but lately it finds its place in all layers of programming in multilayer systems. For the implementation of our system we used only JavaScript.

1) The Database

Recently NoSQL databases gain popularity [9]. This type of database provides a mechanism for storage and retrieval of data using free coherent model unlike the more commonly used relational database. The benefits of this approach include simple design, horizontal scaling and subtle control over the information available. Non-relational database is the best optimized repository containing information of type key-value or document with ID. The purpose is to facilitate the processes of recovery, adding information and introduction of excessive amounts of data and to optimize performance in terms of unintentional delays in the system. Since the data that is processed in this project are supposed to be mainly documents, the most suitable model for non-relational database is the type of document repositories. The central concept behind the Document Repository is the notation for "document". Every document-oriented implementation is different according to the details of the definition for "document", but in all implementations the documents encapsulate and encode data (or information) in any standard formats. The used formats are: XML, YAML, JSON and binary formats like BSON, PDF and Microsoft Office documents (MS Word, Excel, etc.).

Different implementations offer different approach to organizing and grouping of documents:

- Collections;
- Tags;
- Invisible metadata;
- Hierarchy of directories;

The address of a document is represented in the database by a unique key that identifies the document. One of the other characteristic features of document-oriented databases is that besides the simple search by key-document or key-value, which can be used to retrieve the document, databases provide user interface and query language that allow documents to be opened according to their content. In view of our goal to use only JavaScript, we have chosen MongoDB for the development of our system. This database stores the documents in JSON format, which is natural for JavaScript and allows displaying the data in an appropriate form. It is easily readable by developers and is easily parsed by various programming languages. JSON is often used for serialization and transmission of structured information and this implementation is much shorter in code than XML. MongoDB uses JavaScript for querying and administration as well. This makes it particularly suitable for our case.

2) The Serverside and Logic

As for now perhaps the only technology that allows use of JavaScript for writing server-side applications is Node.js. Node.js platform is developed on the Chrome JavaScript environment in order to allow building fast and easily scalable network applications. It uses an event-driven model, which does not block the input and output of the system. This makes it light and efficient and is a suitable choice for applications with very intense real-time data processing and multiple devices handling simultaneously. Another advantage is the large set of additional packages that facilitate programmers work and perform many different tasks: web servers, access to file systems, access to ports, database drivers and much more [10], [11].

Characteristic feature of the event-driven model of Node.js in asynchronous non-blocking mode is that it runs in only one thread, but it still executes parallel and simultaneous operations for requests from many users.

The model that we use for interaction between client and server-side is ReST (Representational State Transfer), a distribution system framework based on Web protocols and technologies [12]. ReST architectural model includes interactions between server and client during the data transfer. The most commonly used module for building ReST architecture is Node.js Express. This is a ready module for building a web server with complete infrastructure for organizing views.

3) User Interface

JavaScript is widely used for building user web interface. There are too many software frameworks that generate HTML document templates. On the first iteration phase we chose the simplest template – Mustache. For our purposes we use mu2Express, as we use Express upgrade to organize our web server and its ReST functionality. This is a simple template for image description. Keywords between braces are replaced by data derived from our database.

VI. Future Work

Testing and evaluation of the system is forthcoming. The development of the modules implementing instruments for QDA will involve longer period of multiple of iterations. And yet the teaching and learning outcomes of the experiment have to be evaluated.

VII. Conclusion

In light of the current requirements for achieving independence from platforms and geographic location of the members of research teams a method for creating cloud-based information system is suggested. A model of the system is developed and technology for its implementation is proposed. The proposed technology is universal and platform-independent and provides easy upgrade and maintenance. The suggested method for developing the project facilitates the continuous development of the system.

References

[1] R. Mary. Making visible the coding process: using qualitative data software in a post-structural study. Issues in Educational Research. Vol.19 (2), 2009, p.142.

[2] S. Merriam. Qualitative research: A guide to design and implementation. San Francisco, CA: Jossey-Bass. 2009.

[3] G. Parkinson, R. Drislane. 2011, Qualitative research. In Online dictionary of the social sciences. http://bitbucket.icaap.org/dict.pl

[4] N. Denzin, Y. Lincoln. (Eds.) Handbook of qualitative research (4th Ed.). Thousand Oaks, CA: Sage. 2011.

[5] P. Nkwi,I. Nyamongo, G. Ryan. Field research into socio-cultural issues: Methodological guidelines. Yaounde, Cameroon, Africa: International Center for Applied Social Sciences, Research, and Training/UNFPA. 2011.

[6] C. Taylor Qualitative Data: at http://statistics.about.com/od/Glossary/g/Qualitative-Data.htm/

[7] Ryan, G., & Bernard, R., 2000, Data management and analysis methods. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 769–802). Thousand Oaks, CA: Sage.

[8] M. Fowler. UML distilled. Third edition. Addison Wesley, 2004.

[9] P.J. Sadalage, M. Fowler. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley, 2012.

[10] M. Kiessling. The Node Beginner Book. Leanpub, 2011.

[11] Node.js v0.10.26 Manual & Documentation. http://Node.js.org/api/

[12] R.T. Fielding. Architectural Styles and the Design of Network-based Software Architectures. Dissertation, Chapter 5. University of California, Irvine. 2000.

Robust Increased Capacity Image Steganographic Scheme

M. Khurrum Rahim Rashid
Department of Electrical Engineering, NUSES FAST, Pakistan

Nadeem Salamat
Department of Mathematics and Statistics
Karakoram International University, Gilgit

Saad Missen, Aqsa Rashid
Department of Computer Science and Information technology, the Islamia University of Bahawalpur, Pakistan

***Abstract*—with the rising tempo of unconventional right to use and hit protection of secret information is of extreme value. With the rising tempo of unconventional right to use and hit, protection of secret information is of extreme value. Steganography is the vital matter in information hiding. Steganography refers to the technology of hiding data into digital media without depiction of any misgiving. Lot of techniques has been projected during past years. In this paper, a new steganography approach for hiding data in digital images is presented with a special feature that it increases the capacity of hiding data in digital images with the least change in images perceptual appearance and statistical properties at too much less level which will be very difficult to detect.**

Keywords—Image steganography, LSB, Security Analysis, Robustness Analysis

I. Introduction

The data that are passing on through internet persistently have possibility of third individual nosy. So there should be some criteria to keep information covert today as the current time is of digital contact. Steganography has established significant attention during the previous few years, in particular after anecdotal news suspected that this tool was used by terrorist. Steganography [2] [3] inquire about to make available a clandestine message control between two parties. There are various steganography [4] [5] technique and medium used for this purpose. These methods are gaining value due to the secret communication over the internet.

LSB substitution [1] and LSB matching method are the oldest methods of steganography. Both of these are applied on least significant bit and only one bit per pixel for grayscale and three bits per pixel for color image can be used for hiding message. During past years lot of techniques of steganography have been projected [4] [5] [8].

In this paper the advantages of LSB matching and LSB substitution are used and capacity is increased with least statistical and perceptual change. The paper is arranged as Section 2 describes Method, Section 3 discusses analysis of cases of change in pixel value, Section 4 compares the Substitution with the proposed method in term of change in pixel value, Section 5 is Compare by Analysis tools and Experimental results and discussion, Section 6 is the conclusion and references are in last section.

II. Method

In the proposed method the two operations are performed on one pixel. First the pixel value is adjusted in such a way that the second least significant bit becomes the message bit and after that substitution is performed on the least significant bit. This technique is applied on both the grayscale and color image.

A. Description of the proposed Method

The proposed system performs operations on second and least significant bit separately to improve the performance of simple substitution on least two bits of pixel. These can be described as:

1) Take two consecutive message bits Xi and X_{i+1}

2) For Xi, if second least significant bit is not same as Xi, then increment and decrement in pixel value in such a way that second least bit becomes the message bit Xi.

3) For the X_{i+1}, if least significant bit in not same as X_{i+1}, then simply the least significant bit of pixel will be replaces by the message bit X_{i+1}.

B. Description of the proposed Method

1-Message bits= X_i X_{i+1}=00
Pixel Decimal=4

00000100	00000100	00000100
(a) Pixel Binary	(b) Pixel binary after SLSB operation	(c) Pixel binary after LSB operation

Transformation of pixel value from (a) to (b) require no change because they are before now the same. Same is the case with transformation of (b) to (c). It is concluded from this that it is a case of no change (NC).

2-Message bits= X_i X_{i+1}=11
Pixel Decimal=4

00000100	00000011	00000011
(a) Pixel Binary	(b) Pixel binary after SLSB operation	(c) Pixel binary after LSB operation

Transformation of pixel value from (a) to (b) require decrement of 1 as message bit and pixel second least significant bit are not same. Transformation of (b) to (c) requires no change because they are before now the same. It is concluded from this that it is a case of change of 1 (C-1).

3-Message bits= $X_i X_{i+1}$=10
Pixel Decimal=4

00000100	00000011	00000010
(a)Pixel Binary	(b) Pixel binary after SLSB operation	(c) Pixel binary after LSB operation

Transformation of pixel value from (a) to (b) require decrement of 1 as message bit and pixel second least significant bit are not same. Transformation of (b) to (c) also creates decrement of 1. So from (a) to (c) total change is 2. It is concluded from this that it is a case of change of 2 (C-2).

4-Message bits= $X_i X_{i+1}$=01
Pixel Decimal=4

00000100	00000100	00000101
(a)Pixel binary	(b)Pixel binary after SLSB operation	(c) Pixel binary after LSB operation

Transformation of pixel value from (a) to (b) require no change. Transformation of (b) to (c) creates a change of 1. So from (a) to (c) total change is 1. It is concluded from this that it is a case of change of 1 (C-1).

C. Formal steps for insertion process:

Input: Cover Image CI and array of bit stream BS
Output: Stego-Image SI

```
1. $P \leftarrow m$
2. $i = 0$
3. $l(n) = length(BS)$
4. $for\ j = 0, \dots, \frac{l(n)}{2}$
     a. $X_i and X_{i+1}\ are\ two\ consective\ bits\ from\ BS$
     b. $SLSBP_j \leftarrow Second\ least\ significant\ bit\ of\ pixel$
     c. $if(X_i \neq SLSBP_j)$
          i. $Adjustment\ of\ pixel\ value\ so\ that\ X_i = SLSB_j$
     d. $end\ if$
     e. $LSBP_j \leftarrow least\ significant\ bit\ of\ pixel\ at\ j\ location;$
     f. $if(X_{i+1} \neq LSBP_j)$
          i. $LSBP_j \leftarrow X_{i+1};$
     g. $end\ if$
     h. $i = i + 2;$
5. $end\ for$
```

Where *l(n)* is the length of message bits, BS is the array of bit stream that contain the message bits. For the grayscale images the loop of step 4 will run for half of the length of the message bits as one interval of loop will hide two message bits in one pixel. X_i and X_{i+1} are the two consecutive message bits from array of bit stream BS.

For color images the loop interval will be reduced as each interval of loop will hide six message bits in one loop interval in one pixel. In sub-step (a) and (b) of step 4, the second least significant bit and least significant bit will be computed for red, green and blue channel for one pixel. Similarly in sub-step c of step 4, six consecutive message bits will be taken from the bit stream for each interval of loop of step 3 and hide two bits in each red, green and blue channel and therefore sub-step (h) of step 4 will increment 6 instead of 2 in each interval of loop.

Adjustment steps are following:

```
1. $if(X_i = 0\ \&\&\ SLSBP_j = 1)$
     a. $if(P_j = 255)$
          i. $P_j \leftarrow P_j - 2;$
     b. $else$
          i. $if(P_j\%2 = 0)$
               1. $P_j \leftarrow P_j - 1;$
          ii. $else$
               1. $P_j \leftarrow P_j + 1;$
          iii. $end\ else$
          iv. $end\ if$
     c. $end\ else$
     d. $end\ if$
2. $else$
     a. $if(P_j = 0)$
          i. $P_j = P_j - 2\ ;$
     b. $else$
          i. $if(P_j\%2 = 0)$
               1. $P_j \leftarrow P_j - 1\ ;$
          ii. $else$
               1. $P_j \leftarrow P_j + 1\ ;$
          iii. $end\ else$
          iv. $end\ if$
     c. $end\ else$
     d. $end\ if$
3. $end\ else$
4. $end\ if$
```

These steps just increment or decrement the pixel value in such a way that the second least significant bit becomes the message bit. Observation of all the 256 shades shows that there is 99.22% possibility that this change will be 1 and for only 0.78% it will be 2.

D. Formal steps for extraction process:

While extraction, the loop will not end as long as at least two bits are collected as message bit from all pixels of the image. This is because the insertion is quite different from the retrieval process. We just recover the two LSBs value of each pixel and translate this to ASCII; the message will be understandable and in readable format up to the point that the message was inserted, and will then come into view as claptrap.

Input: Stego-Image
Output: Message

```
1. $P \leftarrow m$
2. $for\ j = 0, \dots, m$
      a. $LSBP_j \leftarrow$
         $Least\ significant\ bit\ of\ pixel\ at\ j\ location$
      b. $if\left(SLSBP_j == 0\right)$
            i. $BS \leftarrow 0$
      c. $else$
            i. $BS \leftarrow 1$
      d. $end\ else$
      e. $end\ if$
      f. $if\left(LSBP_j == 0\right)$
            i. $BS \leftarrow 0$
      g. $else$
            i. $BS \leftarrow 1$
      h. $end\ else$
      i. $end\ if$
3. $end\ for$
```

In step 1 *m* is the total number of pixels in the image. *SLSBP* and *LSBP* are the least two significant bits of pixel P and *BS* is the bit stream of message bits. After collecting all bits in *BS*, its ASCII conversion will give the message in readable format. If we know the length of the message that was inserted, then the loop will be ended when the length of message is completed and only the message will be retrieved i.e., no gibberish will be seen at the end of the message.

E. Analyses of Cases of Change in Pixel Value

This section makes analysis for all the shades of gray for checking the change in pixel value after the direct substitution and proposed method.

Table 1 shows the cases of change in all the gray shades for substitution method for two bits. Substitution method simply replaces the pixel's least two bits with the two message bits. For example if the two consecutive message bit are 11 and pixel binary is 11111101. Then substitution simply replace 01 with 11, 11111101→11111111.Replacement will be performed if the message bits and pixel's least bits are not same.

In Table I, first column is the gray level; second column is the binary of gray level, third column shows the binary value after substitution of 00, fourth column shows the possibility of C (Change) and NC (No Change) in gray level after substitution of 00. C-1 means the gray level will have change of 1 after substitution, C-2 is a change of 2, C-3 is a change of 3 and NC means that gray level will remain same as before substitution. Similar to substitution of message bits 00, column 5 and 6 shows the possibility of C or NC for the substitution of message bits 11; column 7 and 8 shows the possibility of C or NC for the substitution of message bits 10 and last two columns are for message bit 01.

F. Comprison of Substitution and Proposed Method

From the Table II and IV it is clear that proposed method is better than direct substitution as possibility of C-3 have been decreased from 12.50% to 0.20% by increasing the possibility of C-1 from 37.5% to 49.80%. Change of 3 is a grater change as compare to 1. Change of 1 is invisible to human eye and almost undetectable. So proposed method is better in terms that it decreases the possibility of 3 and increases the possibility of change of 1.

TABLE.I. CHANGE AFTER SIMPLE SUBSTITUTION METHOD

Value	Binary of Value	For Message Bits 00	C/NC 00	For Message Bits 11	C/NC 11	For Message Bits 10	C/NC 10	For Message Bits 01	C/NC 01
0	00000000	00000000	NC	00000011	C-3	00000010	C-2	00000001	C-1
1	00000001	00000000	C-1	00000011	C-2	00000010	C-1	00000001	NC
2	00000010	00000000	C-2	00000011	C-1	00000010	NC	00000001	C-1
3	00000011	00000000	C-3	00000011	NC	00000010	C-1	00000001	C-2
4	00000100	00000100	NC	00000111	C-3	00000110	C-2	00000101	C-1
5	00000101	00000100	C-1	00000111	C-2	00000110	C-1	00000101	NC
6	00000110	00000100	C-2	00000111	C-1	00000110	NC	00000101	C-1
7	00000111	00000100	C-3	00000111	NC	00000110	C-1	00000101	C-2
. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
248	11111000	11111000	NC	11111011	C-3	11111010	C-2	11111001	C-1
249	11111001	11111000	C-1	11111011	C-2	11111010	C-1	11111001	NC
250	11111010	11111000	C-2	11111011	C-1	11111010	NC	11111001	C-1
251	11111011	11111000	C-3	11111011	NC	11111010	C-1	11111001	C-2
252	11111100	11111100	NC	11111111	C-3	11111110	C-2	11111101	C-1
253	11111101	11111100	C-1	11111111	C-2	11111110	C-1	11111101	NC
254	11111110	11111100	C-2	11111111	C-1	11111110	NC	11111101	C-1
255	11111111	11111100	C-3	11111111	NC	11111110	C-1	11111101	C-2

Table II shows the conclusion of Table I.

TABLE.II. CHANGE AFTER SIMPLE SUBSTITUTION METHOD

Total Gray Levels	**NC (No Change)**	**C-1 (Change of 1)**	**C-2 (Change of 1)**	**C-3 (Change of 1)**
256 for 00	64	64	64	64
256 for 11	64	64	64	64
256 for 10	64	128	64	
256 for 01	64	128	64	
Total:				
1024	256	384	256	128
Average:				
100%	25%	37.5%	25%	12.5%

Table III shows the cases of change for proposed method.

TABLE.III. CHANGE AFTER PURPOSED METHOD

Value	**Binary of Value**	**For Message Bits 00**	**C/NC 00**	**For Message Bits 11**	**C/NC 11**	**For Message Bits 10**	**C/NC 10**	**For Message Bits 01**	**C/NC 01**
0	00000000	00000000	NC	00000011	C-3	00000010	C-2	00000001	C-1
1	00000001	00000000	C-1	00000011	C-2	00000010	C-1	00000001	NC
2	00000010	00000000	C-2	00000011	C-1	00000010	NC	00000001	C-1
3	00000011	00000000	C-1	00000011	NC	00000010	C-1	00000001	C-2
4	00000100	00000100	NC	00000111	C-1	00000110	C-2	00000101	C-1
5	00000101	00000100	C-1	00000111	C-2	00000110	C-1	00000101	NC
6	00000110	00000100	C-2	00000111	C-1	00000110	NC	00000101	C-1
7	00000111	00000100	C-1	00000111	NC	00000110	C-1	00000101	C-2
. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
248	11111000	11111000	NC	11111011	C-1	11111010	C-2	11111001	C-1
249	11111001	11111000	C-1	11111011	C-2	11111010	C-1	11111001	NC
250	11111010	11111000	C-2	11111011	C-1	11111010	NC	11111001	C-1
251	11111011	11111000	C-1	11111011	NC	11111010	C-1	11111001	C-2
252	11111100	11111100	NC	11111111	C-1	11111110	C-2	11111101	C-1
253	11111101	11111100	C-1	11111111	C-2	11111110	C-1	11111101	NC
254	11111110	11111100	C-2	11111111	C-1	11111110	NC	11111101	C-1
255	11111111	11111100	C-3	11111111	NC	11111110	C-1	11111101	C-2

CONCLUSION FROM TABLE 3:

Table 4 shows the conclusion of Table 3.

TABLE.IV. CHANGE AFTER PROPOSED METHOD

Total Gray Levels	**NC (No Change)**	**C-1 (Change of 1)**	**C-2 (Change of 1)**	**C-3 (Change of 1)**
256 for 00	64	127	64	1
256 for 11	64	127	64	1
256 for 10	64	128	64	
256 for 01	64	128	64	
Total:				
1024	256	510	256	2
Average:				
100%	25%	49.80%	25%	0.20%

III. COMPARISON BY ANALYSES TOOLS AND EXPERIMENTAL RESULTS

This section discusses the “Comparison and Analysis tools” used to test the proposed method.

A. Comparison and Analyses Tools

The comparison section is further subdivided into two sections. First is named as “Security Analysis” and second is named as “Robustness Analysis”.

B. Security Analyses

Comparing the histograms of cover image and the stego-image gives the clear idea of security. The security examination evaluates the cover image with the stego-image on the basis of histograms of Images. For histogram comparison Correlation, Chi-square, Intersection and Bhattacharya distance [6] are computed between the histogram of cover image and stego-image.

All these comparisons are performed on normalized histogram. The correlation value varies between 1 and -1. Perfect match is 1 and total mismatch is -1. For Chi-square ideal value is 0 and mismatch value is unbound, for intersection 1 is ideal matching value and 0 is mismatched value and Bhattacharya distance gives 0 for the exact match and 1 for mismatch. When these comparison matrices gives ideal values or values that are closer to ideal values then the change in histogram is very least and this is the evidence for Stego-System to be a secure system.

C. Robustness Analyses

Robustness of any method depends on different parameters. In the paper four most important and widely used Image quality measures [7, 9, 10, 11, 12 and 13] namely MSE, PSNR, UIQI and SSIM are computed for comparison.

Mean Square Error computes the perceived error. It is pixel value difference based quality measure. Peak Signal to Noise Ratio [10] is inversely proportional to MSE. Less MSE gives High PSNR which is the proof of the fact that image has good quality.

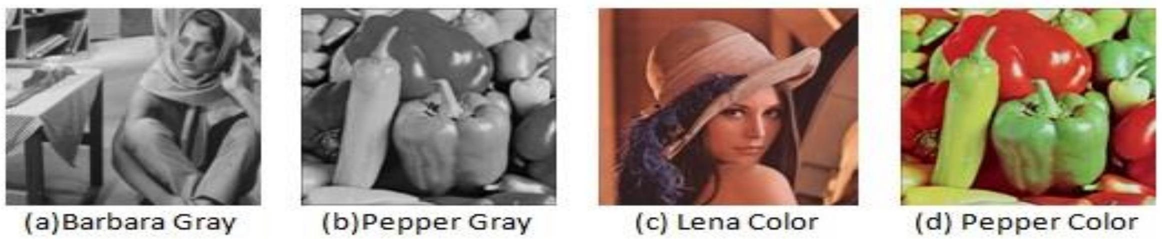
Image Quality Index split the judgment of similarity between Cover Image (CI) and Stego-Image (SI) into three comparisons: Luminance, Contrast and Structural Information.

SSIM estimates "Perceived change in structural information". It computes the similarity between two images of common size. Its mathematical definition is as:

The value of UIQI and SSIM varies between 1 and -1. Closer the highest positive value denotes too much less change in two images and -1 shows totally mismatch. UIQI and SSIM are considered as more consistent and accurate than MSE and PSNR.

IV. Experimental Results

This section presents the experimental results obtained after implementing the proposed method in .NET Framework (C#). A system is designed and implemented in .NET Framework (C#) that shows the functioning of projected Increased Capacity Image Steganography method. The system is named as Robust Increased Capacity Image Steganographic Scheme (RICISS) because of exceptional results of Security Analysis and Robust analysis.



(a)Barbara Gray (b)Pepper Gray (c) Lena Color (d) Pepper Color

Fig. 1. Images used for Test

The proposed method is tested on many standard images. Some from the tested database are shown in Figure 1. (a) is the Barbara grayscale image having dimensions 512x512, (b) is the Pepper grayscale image having dimensions 512x512, (c) is the Lena color Image having dimensions 512x512 and (d) is the Pepper color image having dimensions 225x225.

The section also divided into two subsections. First will give the experimental results for grayscale images and second subsection will give for color images.

A. Grayscale Image

In Barbara Gray Image, Figure 1 (a), different numbers of bits are hidden and results are computed between cover and stego image. Table 5 shows the result of security analysis with 30272 bits of hidden data, 41472 bits of hidden data and 63824 bits of hidden data. Table 6 shows the results of robustness analysis for cover and stego Barbara Gray images.

TABLE.V. Result of Security Analysis for the Barbara Gray Image

Image	Method	30272 Bits	41472 Bits	63824 Bits
Barbara Gray	Correlation	0.99999	0.99998	0.99997
Barbara Gray	Chi-Square	0.00032	0.00044	0.00068
Barbara Gray	Intersection	0.99979	0.99971	0.99956
Barbara Gray	Bhattacharyya	0.00089	0.00104	0.00139

TABLE.VI. Result of Robustness Analysis for the Barbara Gray Image

Image	Robustness Analysis IQM	30272 Bits	41472 Bits	63824 Bits
Barbara Gray	MSE	0.07542	0.10348	0.15983
Barbara Gray	PSNR	59.44863	58.07538	56.18712
Barbara Gray	UIQI	0.99998	0.99998	0.99997
Barbara Gray	MSSIM	0.99998	0.99998	0.99997

In Pepper Gray Image, Figure 1 (b), different numbers of bits are hidden and results are computed between cover and stego image. Table 7 shows the result of security analysis with 24880 bits of hidden data, 41472 bits of hidden data and 55952 bits of hidden data. Table 8 shows the results of robustness analysis for cover and stego Pepper Gray images

TABLE.VII. RESULT OF SECURITY ANALYSIS FOR THE PEPPER GRAY IMAGE

Image	Method	24880 Bits	41472 Bits	55952 Bits
Pepper Gray	Correlation	0.99999	0.99998	0.99997
Pepper Gray	Chi-Square	0.00028	0.00046	0.00062
Pepper Gray	Intersection	0.99993	0.99971	0.99961
Pepper Gray	Bhattacharyya	0.00302	0.00336	0.00352

TABLE.VIII. RESULT OF ROBUSTNESS ANALYSIS FOR THE PEPPER GRAY IMAGE

Image	Robustness Analysis IQM	24880 Bits	41472 Bits	63824 Bits
Pepper Gray	MSE	0.07542	0.10348	0.15983
Pepper Gray	PSNR	59.44863	58.07538	56.18712
Pepper Gray	UIQI	0.99998	0.99998	0.99997
Pepper Gray	MSSIM	0.99998	0.99998	0.99997

B. *Color Image*

This subsection gives the experimental results for the color image. In Lena Color Image, Figure 1 (c), different numbers of bits are hidden and results are computed between cover and stego image. Table 9 shows the result of security analysis and robustness analysis with 66944 bits of hidden data and 81296 bits of hidden data in cover image.

TABLE.IX. RESULT OF SECURITY AND ROBUSTNESS ANALYSIS FOR THE LENA COLOUR IMAGE

Image	Method	66944 Bits	81296 Bits	Method	66944 Bits	81296 Bits
Lena Color	Correlation	0.99999	0.99999	MSE	0.06346	0.07723
Lena Color	Chi-Square	0.00037	0.00044	PSNR	60.09989	59.26520
Lena Color	Intersection	0.99975	0.99969	UIQI	0.99998	0.99997
Lena Color	Bhattacharyya	0.00087	0.00098	MSSIM	0.99998	0.99997

In Pepper Color Image, Figure 1 (d), different numbers of bits are hidden and results are computed between cover and stego image. Table 10 shows the result of security analysis and robustness analysis with 66944 bits of hidden data and 81296 bits of hidden

TABLE.X. RESULT OF SECURITY AND ROBUSTNESS ANALYSIS FOR THE PEPPER COLOUR IMAGE

Image	Method	66944 Bits	81296 Bits	Method	66944 Bits	81896 Bits
Pepper Color	Correlation	0.99992	0.99990	MSE	0.34794	0.41860
Pepper Color	Chi-Square	59.44863	0.00214	PSNR	52.81439	51.99257
Pepper Color	Intersection	0.99889	0.99963	UIQI	0.99993	0.99992
Pepper Color	Bhattacharyya	0.00751	0.00756	MSSIM	0.99993	0.99992

V. CONCLUSIONS

In this paper an increased capacity method of image steganography is presented and implemented for both the grayscale and color images. It makes accessible capacity improvement with least squalor in stego image quality. Experimental outcome be evidence for that the projected technique give good results for security analysis and robustness analysis and thus the projected technique provides the evidence to be strong.

REFERENCES

[1] N.F.Johnson, Sushil Jojadia George Mason University "Exploring Steganography: Seeing the Unseen", (0018-916/98/$10.00©) IEEE, 1998

[2] R.Poornima, R.J.Iswarya, "An Overview of Digital Image Steganography", *International Journal of Computer Science & Engineering Survey* (Vol.4, No 1) 2013

[3] T.Morkel, T.H.P.Eloff, M.S.Olivier, "An Overview of Image Steganography", ICSA Research Group, Department of Computer Science

[4] Jammi Ashok, Y.Raju, S.Munishankaralak, K.Srinivas, Jammi Ashok, “Steganography: An Overview”, et.01./*International Journal of Engineering Science and Technology*, (Vol.2 (10)), 2010, 5985-5992

[5] Shikha Sharda, Sumit Budhiraja , “Image Steganography:A Review”, *International Journal of Emerging Technology and Advance Engineering* (volume 3, Issue 1), January 2013

[6] ASHA, P. NAGABHUSHAN, N. U. BHAJANTRI, “SIMILARITY MEASURES FOR AUTOMATIC DEFECT DETECTION ON PATTERNED TEXTURES”, INTERNATIONAL JOURNAL OF IMAGE PROCESSING AND VISION SCIENCES (IJIPVS) VOLUME-1 ISSUE-1, 2012

[7] Rajkumar Yadav, “Analysis of Various Image Steganography Techniques Based Upon PSNR Metric”, *International Journal of P2P Network Trends and Technology*- (Volume1, Issue2), ISSN: 2249-2615, 2011

[8] M. Pavani, S. Naganjaneyulu, C. Nagaraju, “A Survey on LSB Based Steganography Methods”, *International Journal of Engineering and Computer Science* ISSN: 2319-7242 (Volume 2 Issue 8) Page No. 2464-2467, August, 2013

[9] Ismail Avcibas, Bulent Sankur, Khalid Sayood, “Statistical Evaluation of Image Quality Measure”, *Journal of Electronic Imaging*, 11(2), 206-223(April 2002)

[10] Zhou Wang, *Member,*Hamid R. Sheikh, “Image Quality Assessment: From Error Visibility to Structural Similarity”, *IEEE Transactions On Image Processing*, (VOL. 13, NO. 4), APRIL 2004

[11] Yousra A. Y. Al. Najjar, Dr. D. C. Soong, “Comparison of image quality assessment: PSNR, HVS, UIQI, SSIM”, IJSER, (Vol. 3, Issue8), ISSN2229-5518, August-2012

[12] Amhamed Saffor, Abdul Rahman Ramli, Kwan-Hoong Ng, “A Comparative Study of Image Compression between Jpeg and Wavelet”, *Malaysian Journal of Computer Science*, (Vol. 14 No. 1), pp. 39-45, June 2001

AUTHOR PROFILE:

Muhammad Khurrum Rahim currently is a student of Electrical Engineering BS (EE) in NUSES FAST Islamabad, Pakistan for the session 2013-2017. He has won the competition of English Creative writing in 2007 held in Pano Akil Region, Pakistan by APS&CS. He has one gold and three silver medals in Inter School Mega Competition 2012 and Inter School Mega Competition 2013 in Pano Akil Region, Pakistan by APS&CS. His fields of interest include Robotics, Image Processing, Signal Processing, Circuit theory, Differential and Telecommunication

Aqsa Rashid received her Master’s degree in Computer Sciences (MCS) (Gold Medalist) from Islamia of Bahawalpur, Pakistan in November, 2012. Currently she is a student of MSCS (session Feb, 2013-2015 spring) in Islamia University of Bahawalpur; Pakistan. She is a Visiting Faculty member of Islamia University of Bahawalpur, Department of CS&IT since Nov, 2012 to present. Her fields of interest include Robotics, Digital image Processing, Artificial Intelligence, Data Mining and Web Designing and Development. At present, she is engaged in Image Steganography and Steganalysis.

Saad Missen received his Masters degree (MSCS) in 2005 from University of Management and Technology, Lahore Pakistan. He received his Master of Research degree (M2R) in 2007 from University of Toulouse, France. He received his PhD degree in Computer Sciences in 2011 from University of Toulouse, France. Currently he is working as an Assistant Professor in Dept. of Computer Science & IT, The Islamia University of Bahawalpur, Pakistan. His research interest includes Information Retrieval, Online Social Network Mining, Opinion Mining, Entity Retrieval, Usability Evaluation

Nadeem Slamat received his MPhil degree in Informatics, Mathematics with specialty Image and Calculus from University of La Rochelle, 17000, FRANCE in 2008. He received his PhD degree in Image Understanding from University of La Rochelle, 17000, FRANCE in 2011.He is an HEC Approved PhD Supervisor. Currently he is working as an Assistant Professor in Karakoram International University, Gilgit Pakistan. His research interest includes Fuzzy sets, spatial logic, image understanding, Spatio-temporal Reasoning.

A Hybrid Multi-Tenant Database Schema for Multi-Level Quality of Service

Ahmed I. Saleh
Computers and Systems Department,
Faculty of Engineering,
Mansoura University, Egypt

Mohammed A. Fouad
Information Systems Department,
Faculty of Computer and
Information Sciences,
Mansoura University, Egypt

Mervat Abu-Elkheir
Information Systems Department,
Faculty of Computer and
Information Sciences,
Mansoura University, Egypt

***Abstract*—Software as a Service (SaaS) providers can serve hundreds of thousands of customers using sharable resources to reduce costs. Multi-tenancy architecture allows SaaS providers to run a single application and a database instance, which support multiple tenants with various business needs and priorities. Until now, the database management systems (DBMSs) have not had the notion of multi-tenancy, and they have not been equipped to handle customization or scalability requirements, that are typical in multi-tenant applications. The multi-tenant database performance should adapt to tenants workloads and fit their special requirements. In this paper, we propose a new multi-tenant database schema design approach, that adapts to multi-tenant application requirements, in addition to tenants needs of data security, isolation, queries performance and response time speed. Our proposed methodology provides a trade-off between the performance and the storage space. This proposal caters for the diversity in tenants via defining multi-level quality of service for the different types of tenants, depending on *tenant rate* and *system rate*. The proposal presents a new technique to distribute data in a multi-tenant database horizontally to a set of allotment tables using an *isolation point*, and vertically to a set of extension tables using a *merger point*. Finally, we present a prototype implementation of our method using a real-world case study, showing that the proposed solution can achieve high scalability and increase performance for tenants who need speedy performance and economize storage space for tenants who do not have demanding quality of service.**

Keywords—Multi-Tenancy; Flexible Database Schema Design; Data Customization

I. INTRODUCTION

As the majority of small and medium enterprises are pressured to reduce their expenditure in information technology via cutting down costs spent on buying software licenses and updating the hardware. Therefore, a lot of software vendors turn to the principle of sharing hardware resources, software and services over the Internet among a large number of customers, this environment is called cloud computing, and its customers are called tenants. The cloud software delivery model is called Software as a Service (SaaS), and multi-tenancy is the primary characteristic of SaaS [1], as it allows SaaS vendors to run a single instance of an application and a database to serve multiple tenants with various requirements.

Multi-tenancy increases resource utilization, as well as sharing the same database instance to multiple tenants. However, the more the company shares resources, the more risks it faces because an outage of a shared resource can potentially affect many customers. Shared resources also add to the complexity of the solution [2]. The primary multi-tenant application challenge is how to make the application ready for future tenants' requirements, and enable it to fulfill their interests and business needs, without changing code or database schema and without doing too much work.

Managing data in multi-tenancy database can be divided into three major schemas: *Separated Databases Schema*, which is optimum for security, isolation, and customization, but on the other side, it incurs the highest costs and storage space, moreover, it is hard to maintain a large number of databases; *Separated Tables Schema*, whose cost is low as compared to separated databases, and is suitable for small database applications, where the number of tables per tenant is small, but it has scalability issues since it needs to maintain a large numbers of tables; *Shared Tables Schema*, which achieves the best storage space, the lower costs and good scalability at the expense of poor performance. Each of the aforementioned approaches has special requirements in designing schema process, and selecting the appropriate approach for every application depends on a number of changeable factors, such as the nature of application, the number of participative customers, the number of tables and the importance of data security. Table I shows a brief comparison among the three approaches.

When investigating a pool of potential SaaS customers, we found that there are two types of customers: the first segment, is customers who have high workloads and focused on quality of services requirements, such as performance, security and isolation assurance as fundamental requirements; on the other hand, there are some customers, how have low workloads, are focused on minimizing tenancy costs by reducing the hardware resources required in the shared system as much as possible and sacrificing workload performance.

Based on these tenants' requirements, we propose a flexible multi-tenant database schema, using a set of factors and rates to be used as shift keys between a separated databases schema, a separated tables schema or a shared tables schema in a multi-tenancy database, in order to achieve a good scalability and a high performance with a low storage space, while supporting a large number of tenants with different level of performance.

TABLE I. THE MAIN APPROACHES TO MANAGING DATA IN MULTI-TENANT DATABASE COMPARISON

	Separated Database	Separated Tables	Shared Tables
Data isolation	high	middle	low
Customizability	high	high	low
Scalability	low	middle	high
Maintenance cost	high	high	low
Optimal use of storage space	low	middle	high
General cost	high	middle	low

The proposed multi-tenant database schema satisfy the two segments of customers, and it can be describe as the following:

- Firstly, it is appropriate for tenants who have a high workloads through separating their tables, to confirm the data isolation, high security and reduce joining operation to get the optimal performance in a shared multi-tenant database.
- Secondly, it economizes the costs for tenants who have a low workloads or who do not have demanding quality of service level, through saving storage space, by sharing the same table for these tenants as possible, while preserving a minimum acceptable level of efficiency to get the optimal cost.

The rest of the paper continues with a background review of existing schema mapping techniques in Section II, followed by a full presentation of our multi-tenant database schema technique in Section III. Our case study will be presented in Section IV, and the paper will be concluded in Section V.

II. RELATED WORK

The three major approaches to manage data in multi-tenant databases are summarized in Table I, and discussed in detail in [3] [4]. In addition to, there are several multi-tenant database schema mapping techniques ware presented in [5],[6], the majority of these techniques are derived from the three major approaches of managing data in multi-tenant databases, in order to create a logical isolated database schema for every tenant in multi-tenant databases. We classify multi-tenant database schema mapping techniques in to three segments: techniques using a single approach to represent data in a multi-tenant databases, techniques mixing two or three data isolation approaches, and techniques mixing unstructured data as XML data type and structured data as relation database.

A. *Techniques Deploying a Single Approach*

Some SaaS providers prefer to use a single multi-tenant data storage approach to represent data in multi-tenant databases, to avoid the complexity in database design.

Private Tables' Technique is derived from the separate tables schema, where each tenant has a logical schema consisting of a set of extensions tables. This technique was explicated in [5], and provides a high performance, however neglects the storage space and scalability. It was preferred to use when applications have a few tenants or a few tables [1].

Universal Table' Technique is derived from the shared tables schema, and was referred to as the most flexible technique in [7]. This technique was adopted by SalesForce.com. In the other side, this technique wastes storage space, because of the great use of *null* values, moreover, it harms query performance because it does not support indexes.

Pivot Tables' Technique is derived from the shared tables schema, it was explicated in [1]. This technique eliminates *null* values and supports more flexible extensions at the expense of increasing query processing time for inserting, updating and deleting operations.

The proposed technique in [8], hosted every tenant data in a separated database, and divided tenants' databases into two classes: high performance and low performance. The main important point in this technique is that it measures the tenant workload by transactions per second, and uses this metric to measure overall workloads to get the hardware provisioning policy and the associated scheduling policy.

B. *Techniques Mixing the Separate Tables Schema and the Shared Tables Schema approaches*

There are a lot of techniques ware proposed to representing data in multi-tenant databases such as in [1], [3], [6], [9], [10]. These techniques mixing separate tables and shared tables schemas to achieve balance between scalability, performance, data isolation and storage space with the best cost. However, these techniques achieve some features, at the expense of other features.

M-store' technique was proposed in [3], it saves storage space and prevents null values, at the expense of reconfiguration. Our proposal solves *null* problem relatively to tenant need of performance.

Chunk Table' technique and Chunk Folding' technique ware proposed in [6], these were flexible and reduce the number of tables, at the expense of increasing the queries complexity, because of the huge joining operation. These techniques just focus on vertical partitioning into logical tables 'chunks', however, our proposal adds vertical and horizontal partitioning to save a specific level of performance for every tenant.

An Elastic Schema' technique was proposed in [1], it works on increasing query performance and storing the data in the database as a character large object (CLOB) or binary large object (BLOB) values, to eliminate the impact of BLOB and CLOB values, and divides tables into common tables and virtual extensions. In our proposed technique every attribute has a special *attribute rate* depending on: the type, size in memory and rate of participation between tenants, in order to decide merging or separating this attribute from the base table.

The technique proposed in [9] works on reducing joining operations, by measuring an attribute's importance based on how many tenants share it. The attributes kept in the base table if they have high rate of participation. Unfortunately, this technique ignores the workload for individual tenants, however, in our proposed technique we solve this problem by evaluating the importance of attributes, via collecting the tenants workloads incorporating with the attributes workload.

C. Techniques Mixing an Unstructured Data and a Structured Data

In [10], the technique works on splitting up the common content tables, shared by all tenants, away from the extension tables. The extension tables contain additional information, tenants may need to supply, these tables are stored in XML document. Using XML technique satisfies SaaS providers and tenants' needs, because the extension data can be handle without changing original database schema. However, it down of performance in queries mechanism, because the collecting between unstructured data in XML and structured data in relation database takes more time.

III. Flexible Multi-Tenant Database Schema

A. Flexible Multi-tenant Database Schema Overview

The previous section outlined the common schema-mapping techniques for managing data in multi-tenancy databases. These techniques focus on realizing all tenants' requirements, and ignore the multiple levels of tenants' workloads. However, in multi-tenancy architecture, a single application and database instance should comply with the tenants' needs as a whole, but the tenants on the same server may have multiple requirements with varying qualities [8], according to a set of business factors, such as: information system workload and importance of data security. These factors motivate us to propose a new schema mapping technique to support multiple levels of data isolation, data security and performance for the tenants in the multi-tenant database.

Multi-tenancy in the database tier can be achieved by sharing databases at different levels of isolation, which results in different multi-tenancy database models according to the requirements for each system. There are three main approaches for isolating data in the multi-tenant database: firstly, shared server and separated databases for each tenant, which provide the highest degree of isolation, but it is much more costly; secondly, shared database and separated tables for each tenant, which is lower than the previous approach in isolation issue, but more fair in the cost; thirdly, shared tables approach which provide the worst degree of data isolation, but it is not costly for the majority of tenants and it achieves a good scalability.

Usually, SaaS providers select the appropriate data isolation approach to each application depending on a set of factors that can change over time, such as: the number of tenants, the size of tenants' data, the importance of data isolation, and the desirable security degree. Unless multi-tenant databases are equipped to handle the changes in these factors such as increasing the number of tenants or the size of their data, they will waste a lot of time and effort in reconstructing database architecture, which is illogical and unacceptable, because SaaS applications should be scalable to support the inconstant customers' needs, without affecting the existing tenants' services.

Building multi-tenant applications with incomplete and inconsistent requirements calls for building a flexible multi-tenant database schema to manage the additional tenants information. Building a flexible multi-tenant database schema should take into account all the influential factors of building the multi-tenant system, including the tenants requirements and the SaaS provider expectations, because multi-tenant database schema should scale to multiple levels of tenants, with multiple requirements and multi-quality of service. This motivates us to propose a new dynamic partitioning mechanism to isolate data in a multi-tenant database, that contains a mixed mode of the three main isolation approaches, in order to improve the server utilization and minimize wasted storage space, while keeping appropriate quality of service for each tenant.

The shared tables schema is referred to as "pure multi-tenancy" in [4] [11]. By the same token, we depend mainly on this approach in our system, in addition to using some metrics to separate extensions tables, to get the desired performance for some tenants, who have a high workload. Finally, we may use a separated database in a special cases, to ensure a desired security level for data, so our proposal combines characteristics of shared tables, separated tables and separated databases architectures.

B. The Standard Components to Build a Multi-tenant Database

A flexible multi-tenant database should be based on three components: firstly, *metadata-driven schema architecture*, which allows tenants to add customizable extensions to the common objects or create entirely new customizable objects. The metadata tables save data of each tenant, such as: users data, desirable entities, customized attributes, and reference to the tenants' data in the temporary tables "*generated tables*", generated tables is provisional tables made in order to store the actual data for tenants; secondly, *global unique identifier*, which generates a new unique id for every inserting operation from any user in any table, which will be used to store the data in the generated tables to minimize the volume of metadata columns in the generate tables; thirdly, *runtime table generator* is the most important component, which decides where to save the data in the generated tables or create a new generated table to save the data.

C. Data Partition in a Multi-tenant Database

The Data distribution in multi-tenancy databases should based on the current transactions and the expected transactions on the data. Before we decide what is the appropriate data distribution approach, we ought to answer the following questions:

a) What are the main functions for the system, and the required resources?

b) What are the additional tenants' requirements, and the required quality of services for each tenant?

c) What are the required memory and processing power to access various data type?

d) What is the expected workload for every table?

e) What is the critical point to horizontal data isolation?

f) What is the critical point to vertical data merging?

Note that the requirements of the system are split to the functional requirements that are the services the system should provide and the minimum quality of services to be accepted, and the non-functional requirements are the other services that improve the system properties e.g. performance, response time.

The proposed technique realizes scalability not only by supporting a large number of customers, but also supporting multiple levels of qualities, through supporting multiple levels of data isolation. *Runtime table generator* is responsible for distributing the tenants' data in the *generated tables*, according to set of suggested factors: *system rate*, *tenant rate*, *attribute rate*, and *table rate*. These factors take in account the volume of activity of the enterprise and the system workload, to provide appropriate data isolation approach and level of performance for the enterprises however the volume of activity.

System rate will be determined by the SaaS provider, to define the functional system needs and the minimum cost for the system, by calculating the required storage space, memory, processing power, maintenance procedures and schedules backup.

Tenant rate will be determined for each tenant individually, it is used to measure the non-functional requirements of tenants such as the desirable level of data isolation, the level of data security, the required query performance and number of extra backups. It is used to define the additional cost for every tenant according to his individual non-functional requirements.

Attribute rate is concerned with the growth of the attribute usage, by taking into account the attribute data type and constrains and whether it was an index attribute.

Table rate is a special rate used to determine every generated table workload, by measuring the growth of the data within each table. On the grounds of the experimental study conducted, which we will discuss in details later, we cater for a set of measurements to compute the four rates as in Table II, and present the suggested weights from these measurements.

The multi-tenant solution must fits functional requirements of the system, which represent by '*system rate*', plus the non-functional requirements of the tenant, which represent by '*tenant rate*'. In that manner, general tenant cost equals the minimum system cost '*system rate*' plus the extra tenant cost '*tenant rate*'.

Our proposal system save multimedia objects, like images and secure documents, in a set of separate databases for files and documents. However, each database has a level of security, a data encryption system and a special backup system that refer by the *database rate*. For example, if a tenant has secure documents and hopes to save them in a separated database, he should have a *tenant rate* high enough to allow him to use one of these separated databases, otherwise his tenant rate is not enough to allow him to use this feature.

TABLE II. THE SUGGESTION MEASUREMENTS TO COMPUTE THE SYSTEM RATE, TENANT RATE, ATTRIBUTE RATE AND TABLE RATE

	Measurement Description	**Weight**
System rate	The number of prospective tenants	20 %
	Average of tables per tenant and prospective data	20 %
	The nature of prospective tenants	10 %
	The prospective non-functional requirements of tenants	10 %
	Number of functions and procedures	10 %
	Minimum security level	10 %
	Minimum isolation level	10 %
	Minimum performance level	10 %
Tenant rate	Maximum number of users for this tenant	20 %
	Growth rate of tenant data (average of transactions per period)	20 %
	Special isolation rate	10 %
	Special security rate	10 %
	Special performance rate	10 %
	Number schedule backup	10 %
Attribute rate	Date type	5 %
	Column type(primary key, index)	5 %
	Foreign key	5 %
	Have constraints(unique, check)	5 %
Table rate	Table growth rate = number of rows / number of days from create	10 %
	Growth rate per tenant= number of rows / number of tenants	10 %
	Total current "*general attributes rate*"	80 %

In multi-tenancy environment, SaaS providers wish to reduce wasted storage space and maximize the sharing of resources, although the tenants wish to maximize the isolation and performance qualities, neglecting storage space. Therefore, we propose to use two thresholds the *isolation point* and *merger point*, to detect when to isolate data in multiple tables or to share data in the same table. Data in a multi-tenancy database will be partitioned horizontally to a set of allotment tables according to *isolation point*, and vertically to a set of extension tables according to *merger point*.

Isolation point is a threshold point that determines the necessity of horizontal partition data, in order to constrict the tables workloads with a concrete point, and it was assigned by the system variables using a several methods according to the nature of application.

The following equations used to calculate the *Isolation point* depending on the isolation factor only, where IP is the isolation point, IR is the isolation rate, SR is system rate, TR is tenant rate, AR is attribute rate and GTR as generate table rate. By the same token, $\mathrm{IR}_{\text{current System}}$ is the current isolation degree of the system, to be accepted by the customers, and it detect by the system provider depending on the nature of the system, $\mathrm{IR}_{\text{average of tanants}}$ is the average of special isolation rates for the tenants who sharing the system, $\mathrm{IR}_{\text{general}}$ the average degree of data isolation rate for both the tenants and the system, note that $\mathrm{IR}_{\text{general}}$ must be less than or equals 1.

$$IR_{total} = IR_{current\ System} + IR_{average\ of\ tanants} \quad (1)$$

$$IR_{maximum} = IR_{maximum\ system} + IR_{maximum\ tanants} \quad (2)$$

$$IR_{general} = \frac{IR_{total}}{IR_{maximum}} \quad (3)$$

$$IP = \left(1 - IR_{general}\right) * \ SR_{maximum} \quad (4)$$

In (5), $AR_{general}$ is a general attribute rate, that calculate a new attribute expected workload, and the required resources by a special tenant.

$$AR_{general} = TR * \ AR \quad (5)$$

Equation (1), (2), (3) and (4) calculate the isolation point, that support multi-levels of data isolation feature. In the same manner we can rebuild these equations where IR replace with the security rate to support multi-levels of data security.

Merger point is a threshold point, that responsible for vertical partition data, in order to reduce the volume of joining operations, it was assigned by the SaaS provider according to a set of factors, such as: the available storage space, the space will be allocated to each tenant, the number of tables will be allocated for each tenant, the expected extra customization attributes for each tenant, and availability to save a null values.

In the proposed system, when a tenant needs to create a custom table or alter an existing table, the *runtime table generator* searches for the best isolation approach by selecting the ideal *generated tables* to save the data in them. The algorithm in "Pseudocode 1" explain how the *runtime table generator* add a new custom table 'ACT' to a tenant 'test'. The algorithm starts by saving the data of a new customized table in the metadata-driven schema tables (line 3), then selecting or creating the appropriate generated tables, to store the tenant data and refer to it in the metadata-driven schema tables. In (lines 4-6) the algorithm check if the new *general attribute rate* is more than or equal the '*isolation point*', then creating a new generated table with this schema and returns its identifier as a reference. Otherwise the algorithm searches in all generated tables, and if it finds a generated table has the new customized table schema and its table rate plus the new general attribute rate are less than or equal to the *isolation point*, then returning this generated table Identifier as a reference (lines 7-12). In the failure case, the algorithm attempts to find all the tenants who need this schema, and if the summation of their *tenant rates* plus *tenant rate* of 'test' is more than or equal to the *isolation point*, a new generated table is created and its identifier is returned (lines 14-19). The last case, a new customized table is divided to a set of sub tables, and the algorithm determines whether this schema is available or it needs to create a new generated tables and return the tables references (lines 20-23).

Pseudocode 1 Creating a new Customized table 'ACT' to tenant 'test'.

```
1:  IP  ← Isolation Point
2:  MT ← Meta-data Table
3:      insert ACT into MT
4:      if ( TenantRate (test) >= IP )  then
5:      create (a new generate table with ACT schema)
6:      return  ( the new generate table ID )
7:      else    for each (GT in Generated Tables)
8:                      GT ← current generated table
9:                      if (Schema(GT)=Schema(ACT))
                        and (TableRate(GT)+TableRate(ACT)<=IP)
10:                     then return  ( the current generated table ID )
11:                     end if
12:             end for
13:     end if
14:  X ← distributed generated tables require ACT schema
15:  if (TableRate (X) + TableRate (ACT) >= IP)  then
16:         create (a new generate table  with  ACT schema)
17:         update MT replace with the new generate table ID where X
18:         move data in X into  the new generate table
19:         return  ( the new generate table ID )
20:         else   for each( sub table  in ACT )
21:                [: : :] /* Symmetric to lines 7 to 19 */
22:                [: : :] /* Symmetric to lines 5 to 6   */
23:                end for
24:  end if
```

D. Analyzing the performance of the proposed multi-tenant database schema-mapping technique

The traditional DBMSs usually consists of 4 basic operations: selection, insertion, deleting and updating operation, and each operation is a collection of an I/O processes in the DBMS. The final target of our solution is twofold: firstly, the system aim at reducing the number of the I/O processes by saving the tenant data in one table. This means that a tenant inserts or updates a row to one physical source table, which means a small response time and a high throughput rate. This serves the tenants whose *tenant rate* is high enough, or high number of tenants with the same schema to allow the merge of their data in one table, which is expected to provide a high quality of service while wasting a lot of storage space; secondly, the system aim at reducing the number of generated tables by maximizing the sharing tables between tenants, by dividing the tenants entities vertically to a set of extension tables in order to eliminate the null values and economize on the storage space for tenants who do not have demanding quality of service, where a single operation can be divided into several I/O processes to the corresponding physical tables, which means a high response time and a lower throughput rate. This approach provides a relatively low quality of service while economizing on storage space.

IV. Experimental Study

In this section, we conduct a case study for a customer relationship management system in multiple hotels to evaluate our proposal. In the beginning, the hotels organizations are divided into multiple levels categories, as illustrated in Table III. All hotels have a common entities such as: room, travel agent and guest entity, and a common procedures such as: reservation operation. However, multiple hotels have different services types, citizenships of clients, and various quality of services. As a result, the common entities of multiple hotels will vary in attributes such as: table of customers. Table IV illustrates a sample of the variety on schema for an entity 'customer' for seven hotels.

The case study consists a sample of customers from “tenant1” to “tenant7”, representing the different segments of hotels. According to the column "Tenant Rate" in Table IV, there are a small enterprises such as “tenant1”and “tenant2”, who have a small workload and do not have demanding quality of service, the system ought to economize storage space for these tenants. However, tenants as “tenant7” and “tenant6” reject to share their tables with others.

In Table V, we compare the performance of our proposed approach with the three main approaches of data isolation in multi-tenancy database. We realize the proposed solution by applying the following steps:

TABLE III. THE MAIN CATEGORIES OF THE HOTELS COMPARISON

Hotel Class	Percentage of Market	Avg. of Users	Avg. of Reservations	Avg. of Transactions	Security Need	Variability need
three stars	80%	2	224	20	Low	Low
four stars	10%	3	689	176	Medium	High
five stars standard	5%	6	734	477	High	High
five stars deluxe	5%	10	1518	158	V. High	V. High

TABLE IV. VARIANT DEFINITIONS OF 'CUSTOMER' ENTITY FOR VARIANT TENANTS

Tenant name	Hotel Level	Tenant Rate	Attribute 1	Data Type	Attribute 2	Data Type	Attribute 3	Data Type	Attribute 4	Data Type
TENANT7	*****D	80	Id	*Int*	Name	*Char 150*	Web Site	*Char 100*	Email	*Char 100*
TENANT6	*****S	69	Id	*Int*	English Name	*Char 150*	Arabic Name	*N Char 200*	Phone	*Char 20*
TENANT5	*****S	49	Id	*Int*	Name	*Char 100*	Address	*Char 500*		
TENANT4	****	34	Id	*Int*	Name	*Char 200*	Tel	*Char 20*		
TENANT3	****	28	Id	*Int*	Short Name	*Char 50*	Full Name	*Char 200*	Site	*Char 100*
TENANT2	***	20	Id	*Int*	Name	*Char 100*	Email	*Char 200*		
TENANT1	***	9	Id	*Int*	Name	*Char 50*	supervisor	*Char 200*		

a) Table VI, we studied the functional requirements for the prospective tenants and the critical system needs to get the system rate.

b) Table VII, we got the tenant rate for each tenant, through studying the non-functional requirements for them, and calculated the extra costs for each tenants.

c) Using equation (4) to get the isolation point through depending on isolation rate.

d) Table VIII illustrate the generated tables GT_i where $i = 1,2,\ldots 9$. schemas, and how to use these to represent the entity 'customer' for all tenants.

A. Implementation with the current isolation rate

In the current case, the system provider set the *system isolation rate* = 5, the minimum= 0 and maximum = 10. Note that the maximum *system rate* equal 100.

$$\text{Total isolation rate } = \frac{5}{10} + \frac{31}{70} = 0.5 + 0.44 = 0.94$$

$$Maximum\ isolation\ rate = \frac{10}{10} + \frac{70}{70} = 1 + 1 = 2$$

$$General\ isolation\ rate = 0.94/2 = 0.47$$

$$Isolation\ point = (1 - .47) * 100 = 53$$

In our case study, isolation point equal 53, so that tenants such as: “tenant$_6$”, “tenant$_7$” who have tenant rate more than or equal 53 will be in separated tables, but the other tenants will be in a shared tables according to their tenant rates.

B. Implementation the minimum system isolation rate

In this case, data isolation quality is not necessary in the system, so the system in common is opt for working like a *shared tables schema*, so it resulted *isolation point* =78. In the final analysis, only “tenant$_7$” who have a *tenant rate* >= 78 will be in a separated tables, and other tenants will be in *a shared table schema*. In conclusion, this case work like a shred tables schema and neglect the performance for most tenants. Finally, it use less number of tables and less storage space.

TABLE V. CHARACTERISTIC OF REALIZATION THE FOUR SCHEMAS

Solution	Tenant cost	customizable	Space requirements	Handle db size	Prospective tenants
Separate databases	V. High	V. High	V. High	Low	5 star Hotels
Separate schemas	Medium	High	Medium	High	4 star and part of 3 star
Shared Table	Low	High	Low	V. High	3 star
Propose Solution	made to order	V. High	made to order	High	All Hotels

$$General\ isolation\ rate = (0 + .44) / 2 = 0.22$$

$$Isolation\ point = (1 - 0.22) * 100 = 78$$

C. Implementation the maximum system isolation rate

In this case, it is opt for working like a *separated tables schema*, where all tenants having a *tenant rate* >= 28 will be in a separated tables. In conclusion, this case use more tables and neglect the storage space size and the number of tables.

TABLE VI. ILLUSTRATE SYSTEM RATE MEASUREMENT

Measurement Description	Weights rate
The number prospective tenants	8
Average number of table per tenant	5
the nature of prospective tenants	2
the needs of prospective tenants	3
Number of functions and procedures	3
Security Rate	4
Isolation Rate	5
Performance Rate	4
the system rate	**51**

TABLE VII. ILLUSTRATE THE TENANTS' RATES IN THE CASE STUDY

	No of users	Growth rate	Isolation Rate	Security Rate	Performance rate	No of Backup	tenant rate
Max. Value	**20**	**20**	**10**	**10**	**10**	**10**	**80**
Tenant 1	2	3	1	1	1	1	9
Tenant 2	6	6	1	1	4	2	20
Tenant 3	8	8	2	3	5	2	28
Tenant 4	8	10	5	2	6	3	34
Tenant 5	15	16	5	5	5	3	49
Tenant 6	18	20	7	8	8	8	69
Tenant 7	20	20	10	10	10	10	80
	AVG		31/70	30/70	39/70		

TABLE VIII. ENTITY 'CUSTOMER' SCHEMA IN THE GENERATED TABLES

Table name	Column 1	Column 2	Column 3	Column 4	Column 5	Shared by tenants	Table rate	Available points
GT1	UID	INT	char 150	char 100	char 100	T7	80	0
GT2	UID	INT	char 150	N char 200	char 20	T6	69	0
GT3	UID	INT				T5	49	4
GT4	UID	char 100				T5	49	4
GT5	UID	char 500				T5	49	4
GT6	UID	INT	char 200	char 100		T2,T3	48	5
GT7	UID	char 50				T1,T3	37	16
GT8	UID	INT	char 200			T1,T4	43	10
GT9	UID	char 20				T4	34	19

In Table VIII, the generated tables start with the *Global Unique Identification* column, that is represented by (UID) column, and then list the attributes which the generated tables ware contained. The seventh column 'Shared by tenants' shows who the tenant are used this generated table, such as GT1 is specified for "tenant$_7$". The eighth column is the table rate equal summation of its *tenants' rates*. finally the last column is the available points is *table rate* subtracted from *isolation point*, the tables have available point more than zero can be shared with tenants have tenant rate less than or equal this rate.

In the final analysis, the three implementation cases achieved the required theory, because the system had a multiple levels of data isolation. Fig. 1 illustrates the variety of data isolation approaches in the system, where vertical axis refer to the percentage of tenants and horizontal axis refer to the $\text{System Isolation Rate}_{=}\ i\ where\ i = 0, 0.1, 0.2, \dots 1$. Furthermore, Fig. 2 illustrates how the needs of tenants can automatically affect the isolation rate. Decreasing the isolation rate mean increasing the separated tables or the storage space.

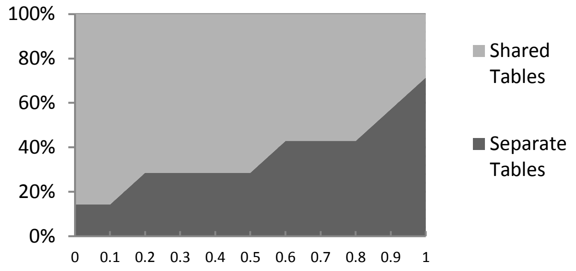


Fig. 1. Example of variation of data isolation schemas in a single system

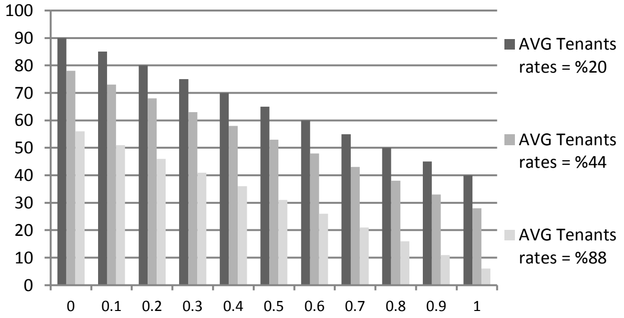


Fig. 2. The effect of change the avarage of tenants rates on isolation rate

V. CONCLUSION

In this paper, we trade-off between the performance and the storage space, in the major approaches of managing data in multi-tenant databases, as well as highlighted the standard components to build a customizable multi-tenant database schema. This paper proposes a new multi-tenant database schema-mapping technique, contain a mixed mode of the multi-tenancy data isolation approaches, in order to improve the database server utilization and minimize the wasted storage space, while keeping the appropriate quality of service for each tenant, by selecting the effective way to part the data in multi-tenant databases horizontally by the isolation point and vertically by the merger point.

REFERENCES

[1] H. Yaish, M. Goyal, and G. Feuerlicht, “An Elastic Multi-tenant Database Schema for Software as a Service,” 2011 IEEE Ninth Int. Conf. Dependable, Auton. Secur. Comput., pp. 737–743, Dec. 2011.

[2] R. F. Chong, “Designing a database for multi-tenancy on the cloud Considerations for SaaS vendors,” pp. 1–12, 2012.

[3] M. H. M. Hui, D. J. D. Jiang, G. L. G. Li, and Y. Z. Y. Zhou, “Supporting Database Applications as a Service,” 2009 IEEE 25th Int. Conf. Data Eng., pp. 832–843, Mar. 2009.

[4] M. N. A. Khan, A. Shahid, and S. Shafqat, “Implementing a Storage Pattern in the OR Mapping Framework,” Int. J. Grid Distrib. Comput., vol. 6, no. 5, pp. 29–38, Oct. 2013.

[5] S. Aulbach, M. Seibold, and S. a P. Ag, “A Comparison of Flexible Schemas for Software as a Service,” Proc. 35th SIGMOD Int. Conf. Manag. data, pp. 881–888, 2009.

[6] S. Aulbach, T. Grust, D. Jacobs, A. Kemper, and J. Rittinger, “Multi-tenant databases for software as a service: schema-mapping techniques,” Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 1195–1206, 2008.

[7] Liao, K. Chen, and J. Chen, “Modularizing Tenant-Specific Schema Customization in SaaS Applications,” AOAsia ’13 Proc. 8th Int. Work. Adv. Modul. Tech., pp. 9–11, 2013.

[8] W. Lang, S. Shankar, J. M. Patel, and A. Kalhan, “Towards Multi-Tenant Performance SLOs,” Data Eng. (ICDE), 2012 IEEE 28th Int. Conf., pp. 702 – 713, 2012.

[9] J.. Ni, G.. Li, J.. Zhang, L.. Li, and J.. Feng, “Adapt: Adaptive database schema design for multi-tenant applications,” ACM Int. Conf. Proceeding Ser., pp. 2199–2203, 2012.

[10] S. Foping, I. M. Dokas, J. Feehan, and S. Imran, “A new hybrid schema-sharing technique for multitenant applications,” 2009 Fourth Int. Conf. Digit. Inf. Manag., 2009.

[11] Bezemer and A. Zaidman, “Challenges of Reengineering into Multi-Tenant SaaS Applications,” 2010.

A Novel Approach for Bioinformatics Workflow Discovery

Walaa Nagy
Information Systems Dept.
Faculty of Computers and Information
Cairo University.
Egypt

Hoda M. O. Mokhtar
Information Systems Dept.
Faculty of Computers and Information
Cairo University.
Egypt

***Abstract*—Workflow systems are typical fit for in the explorative research of bioinformaticians. These systems can help bioinformaticians to design and run their experiments and to automatically capture and store the data generated at runtime. On the other hand, Web services are increasingly used as the preferred method for accessing and processing the information coming from the diverse life science sources. In this work we provide an efficient approach for creating bioinformatic workflow for all-service architecture systems (i.e., all system components are services). This architecture style simplifies the user interaction with workflow systems and facilitates both the change of individual components, and the addition of new components to adopt to other workflow tasks if required. We finally present a case study for the bioinformatics domain to elaborate the applicability of our proposed approach.**

***Index Terms*—Web services; In-silico Workflows; Quality of Services (QoS); Web services for bioinformatics; Bioinformatics services.**

I. Introduction

Due to the large number of available web services, the sheer complexity of data, and the frequent lack of documentation, discovering the most appropriate web service for a given task is a user challenge. In addition, in current practice there is a disparity between bioinformatics workflow conceptualization and specification. We believe that this disparity can be resolved using semantic Web. Semantic Web technologies can be used to translate from the specification of analytical functions or final state to executable workflows. Such workflow abstraction can consequently result in a unification of workflow conceptualization and specification. Nevertheless, such a level of abstraction will be useful for both novice and expert users, and will provide a means to easily and efficiently create workflows for other life sciences researches. Many life science tools are currently available as web services. The use of a workflow system to orchestrate these services and create in-silico experiments seems to be a logical approach. Thus, the methodology for creating bioinformatics workflow depends heavily on the knoweledge that the researcher has about the tools that he will use. In our research the user was enabled to emphasize the desired final state instead of providing the details of the process that he wants to perform or the relationships between his/her tools, which corresponds more closely to the objective of the research, provides further abstraction and reduces the workflow conceptually to a single functional operation as follows:

construct phylogenetic tree to a reference sequence.

This statement ”Construct phylogenetic tree” subsume the steps required to generate phylogenetic tree [1] which are:
Step 1: Choosing an appropriate markers for the phylogenetic analysis.
Step 2:Performing multiple sequence alignments.
Step 3:Selection of an evolutionary model.
Step 4:Phylogenetic tree reconstruction.
Step 5:Evaluation of the phylogenetic tree.

Currently, to the best of our knowledge, there is no guide to assist users in the Web services discovery process. However, there is a significant progress that has been made towards building integration platforms that utilize online Web services interfaces to support bioinformatics analyses. Taverna [2] is such a platform that is widely used by the bioinformatics community, whereas Kepler [3] and Triana [4] are two popular platforms in the wider scientific community. Recently, there is a trend to extend such platforms to support semantic Web services, as more semantically linked data repositories are now available. Semantic annotation provides richer information than Web service description alone and can be used for automatic reasoning when they conform to certain ontology.

Other approaches focus on the discovery of Web services that are annotated with a specific vocabulary. This is the case of the myGrid[1] project, whose aim is to provide a controlled vocabulary to make annotations. BioCatalogue [5] is a life science Web service registry with 1180 registered Web services that are meant to be annotated using life sciences ontology. Nevertheless, another issue to be taken into account is that, in many cases, multiple services provide very similar functionality (a particularly insidious example is the multitude of services providing variants of alignments of genes and proteins). In such cases, the user has to decide which service is the most appropriate based on diverse quality criteria (availability, coverage of the domain of interest, etc.). To

[1] http://www.mygrid.org.uk

address this problem, assessment techniques must be applied to provide the user with some information about the quality and the functionality of each service [6].

In this research a workflow representation was suggested that can solve the disparity between how life scientists conceptualize their workflows and how the workflows are specified in practice. The resolution is through a description of workflows through specification of primary analytical operations or desired analysis final state that enables the design of high-level workflow models. These high-level workflow descriptions are closer to the conceptual models that life scientists have about experiments. Consequently, the cognitive load on those scientists has been reduced [7]. In our work we use the semantic Web technology to translate from final state functions to an executable workflow.

The proposed idea depends on the idea of standardization of Web service interfaces. This standardization enables the language to be later used in life science workflow systems to describe and implement classes of tasks. For example, the proposed workflow representation can help to provide a standardized interface for the famous biological sequence alignment tasks in a workflow, then; the abstract alignment task encompasses all alignment resources, including (a) synchronous Blast services, and (b) Blast services. There are several key design rationales of our language that can be summarized as follows:

1) Our proposed services selection for in-silico workflow enables the life scientist to think in terms of the operation he wants to perform instead of the services he needs to use. Thus our solution reduces the gap that currently exists between the level at which a life scientist thinks about the life science problem and the actual workflow implementation.
2) The proposed workflow has all-service architecture, i.e., all system components are services. This allows easy change of individual components and addition of new ones to adapt to other workflow tasks if required.
3) The whole system is based on a semantic Web service framework (WSMO [8]). As a result, all system components provide native semantic support.
4) In our proposed workflow services selection method only essential services are available and automatic service selection is done. .
5) the proposed approach select the appropriate Web services that can execute the task dynamically at run time instead of hard coding the services that can execute the task during the design of the workflow. This feature in turn overcomes the issue of dead or unavailable services.
6) Finally, the user provided with the best available Web services that performs his request based on both quality of services criteria, and his/her stored profile values. This feature considers the fact that currently there exist potentially large number of services from different service providers that can functionality overlap we believe that the non-functional properties of services, especially those related to Quality of Service (QoS), such as reliability, performance, and sensitivity, should be considered when several services provide similar function or information. A service selection algorithm is used to help users identify the "best" service [12].

The rest of the paper is organized as follows: Section II, presents background, related work, and motivates the problem. Section III, discussed the workflow discovery problem and solution.Section IV explain the worklet architecture. Section V, presents a use case that are used to evaluate the proposed solution. Finally, section VI concludes and outlines directions for future work.

II. RELATED WORKS

A. Current Workflow Design Systems.

There have been a large number of workflow design and execution engines that supports In silico biological study. Taverna [2] which is a workflow construction environment and execution engine designed to support in-silico biological study. It provides access to a large collection of data sources and analysis tools, many of which are accessed through a Web service interface. Taverna is designed as a do-it-all environment, which can be overwhelming for biologists with limited computing background. While Taverna has a plug-in architecture that allows addition of new functions, any changes to the core system components are non-trivial.

In [9] a prototype solution for the service selection problem in the life-sciences is proposed. The solution uses the Moby environment that uses lazy breadth-first search over an implicit graph of the service space. Also, the presented method is able to highly rank desirable services at interactive speeds. Although the solution improves the service selection process so that the user is presented only with a small set of the most salient services to the time and effort required to build workflows, the solution purposely uses a minimal amount of semantic information (data-type matching only), and does not consider the un-experienced user who is not familiar with workflow usage.

B. Semantics and quality-aware Service (QoS) in Current Workflow Design Systems.

Some workflow design systems have incorporated semantic component to support workflows execution. A bioinformatics semantic workflow is the workflow specification of a complex scientific task in terms of the biological analysis objectives and the semantic characteristics of the computational tools needed for the analysis [10] without delving into the computational details of the tools. The semantic workflow is at the level of biological concepts, that are much closer to the scientific research. Although Taverna has semantics to describe workflow activities, it is limited to supporting the discovery of proper services during the design of a workflow [9].

In [10], the author presented ***Sesame*** a semantic bioinformatics workflow design system with new ontology for bioinformatics tools/services. Sesame allows the biologists to perform their analyses using terms that they are familiar with. After designing the semantic workflow, ***Sesame*** ask the user to instantiate it by associating each analyses entity with the instances of bioinformatics tools/services and data.

Although Sesame free the biologists from the necessity of learning the details of the computational aspects of the bioinformatics tools. yet, it still requires its users to have knowledge in scripting tools/ services, and algorithms that they will use.

Also, ***Sesame*** can only perform simple instantiation cases and for each analyses entity Sesame ask the user to select one instance of bioinformatics tools/services. Then, the user specifies the parameters and input data for the selected tool/service which overload the unexpirencied user with alot of work that he/she may not understand.

The work in [11] presents a new way for describing workflow templates and instances. All the used constituents are semantic objects that are described with properties and workflow level constraints. Once a workflow template is created and validated by an experienced user, it is easy for more junior scientists to create sophisticated analyses simply by specifying input data for pre-defined templates. The system ensures that the input data specified is appropriate given the definitions in the workflow template, and automatically generates a workflow instance that can be mapped to execution resources.

Despite all those efforts, in this work a better approach was presented that ensures the existence of all the resources that can perform the task rather than having a template that may contain a link to a dead service or an unavailable resource. In addition, the proposed approach also consider the QoS information in selecting the services that best execute and meet the user requirements.

The authors in [12] have utilized several semantic technologies to identify the scientists intent, and then to facilitate the control of workflow execution and enrichment of workflow provenance. While the integration of semantic components with the original system improved the usability of the workflow system, users still have to be familiar with those workflow design systems in order to accomplish a data analysis task.

Zhang et al. proposed a two-step approach to automatically transform geospatial procession conceptual workflow to Kepler workflow [3]. However, their transformation is limited to only one geographical information system package.

For a more complete selection of services in addition to the syntactic and the semantic dimension in building the workflow the quality of resources are also important factors for making adequate service choice.

In [13], an optimization algorithm is proposed to efficiently access Web services. The algorithm takes as input the classical database SPJ like queries over Web services. It uses a cost model to arrange Web services in a query, and computes a pipelined execution plan with minimal total query running time. In addition, in [14] quality-aware service optimization techniques have been studied. These approaches rely on the computation of a predefined objective function, and the users need to assign numeric weight to specify their preferences if multiple quality parameters are involved.

On the other hand, in our proposed approach we learn those weights dynamically from the user profile instead of asking the user about the weights of QoS criteria which could lead to missing the user desired services due to the user inability to precisely specify the weight of each QoS parameter if multiple quality parameters are involved. Nevertheless, our proposed approach follows the work presented in Adams et al. [15]. In this work the authors introduced a technique to design abstract workflows to describe the treatment processes of patients. The central component in their approach is worklets. A *worklet* is a small workflow that covers all the actions required to perform a higher level task (or step). The main advantage of worklets is that it black boxes implementation details from a workflow designer point of view. The approach supports flexibility and evolution in workflows through the support of flexible work practices, based not on proprietary frameworks, but on accepted ideas of how people actually work. Our approach although borrows the worklet concept, differs in the way we build and select the worklet assigned for each workflow task. In addition, we present a new application direction for workflows through employing worklets that target bio-informaticians. In addition, we employ worklets to achieve an efficient Web service selection interface. The remaining of the paper elaborates our approach in using worklets to provide a standard language for in-silico service selection.

III. Workflow Discovery

Workflow discovery in our work is responsible for finding available workflows for the specified functionality. Workflows are discovered using information on their specification, the desired functionality, and available services. For simplicity, we provide an example of workflow discovery using only the desired functionality (e.g., phylogenetic analysis).

Each task of a workflow is linked to an extensible repertoire of actions. In this work, we present those repertoire-member actions as “worklets”. In effect, we borrow the definition of a worklet from [15] where a worklet is defined as a workflow that handles one specific task in a larger, composite workflow. The use of worklets enables the design of abstract, reusable workflows that can be used for different cases with a similar high level structure. Finally, a sequence of worklets is chained to form an entire workflow process.

At the same time, a repertoire for each workflow task is dynamically constructed as different approaches and methods for completing those tasks are continuously developed. The input variables of the original task are mapped to the input variables of the selected service on the worklet, and then the worklet is launched. When the worklet is completed, its output variables are mapped back to the output variables of the original work process.

Our workflow execution use the same model of execution presented in [16]. In [16] the workflow is composed of a service that can execute the required task but our workflow is composed of worklets that enable the design of abstract, reusable workflows that can be used for different cases with a similar high level structure. The execution of a workflow depends on a utility service such as *satisfyPrecondition*, this service is called at the beginning of each process model for worklets with preconditions. As a result, the prerequisite worklet are executed first, in order to satisfy the preconditions; once all preconditions are met, the initial service is finally executed. The precondition dependency check is recursively applied until a service with no preconditions is reached.

In this way, each service prerequisite will help compose an executable workflow. Subsequently, the workflow will be composed backward and at runtime reversed for proper execution. Figure 1 shows this method.

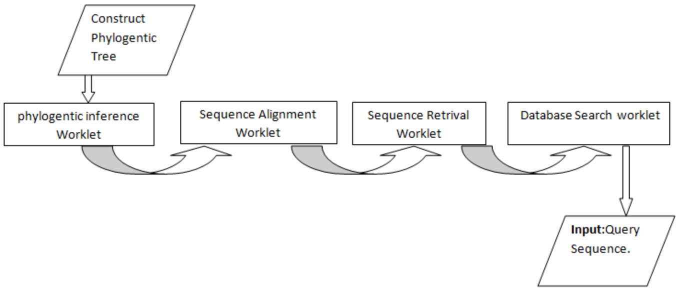


Figure 1: Workflow Backwards Composition

A worklet in our work is handled like a normal Web services. Semantic services are defined using OWL-S [17]. A semantic worklet service description is constructed with a precondition that reflects the prerequisites needed in order for the service to successfully complete; this precondition is sufficient to construct workflows for the desired functionality. We have considered a way to define worklets preconditions. For example, a phylogenetic analysis worklet named “Phylogenetic Inference Worklet, This service requires a multiple sequence alignment as an input and can be described as follows:

Worklet: Phylogenetic Inference.
Precondition: Multiple sequence alignment.
Input: Optional multiple sequence/analysis parameters.
Output: Phylogenetic Tree.

As we are searching for life sciences Web services, we apply the taxonomy of categories used by BioCatalogue [5] in order to classify the worklet. BioCatalogue is a registry of curated Life Science Web Services. The aim of BioCatalogue is to provide an easy way for scientists to discover Web services of interest. BioCatalogue has a shallow taxonomy of Web services categories, and most of the registered Web services have at least one category. So, we use this taxonomy to classify the worklets with the aim of using those categories in the discovery of the Web services that are suitable for a given worklet function.

Example 3.1: Consider a worklet ”Sequence Alignment” building problem. In this worklet we collect all services that can perform sequence alignment like (a) synchronous Blast services, and (b) Blast services. Hence, the user simply specifies that a Sequence Alignment is desired. Then, we select the worklet named Sequence Alignment to execute his request. At the lowest level, expert users can specify all the analysis parameters for every step in each analysis that constitutes the workflow.

Web services in BioCatalogue have four main different types of annotations: descriptions, operations, tags, and categories. Each worklet will contain a set of comparable services in function, the comparability is derived from the bioinformatics ontology annotation that the service has. For instance, two services are comparable if they have the same function annotation, in the following section we will provide the details of how we discover those services.

IV. Worklet Architecture

A worklet is basically nothing more than a workflow specification that has been designed to perform one part of a larger parent specifications. However, a worklet differs from a decomposition or sub-net in that it is dynamically assigned to perform a particular task at runtime, while sub-nets are statically assigned at design time. So, rather than being forced to define all specification of the task such as algorithms to be used and the service during the design time of a workflow, the worklet service allows the definition of a much simpler specification that will evolve dynamically as more worklets are added to the repertoire for particular tasks.

In addition, as we aim to support flexible system architecture, we therefore implement the worklet components as Web services as well. Such implementation approach enables us to easily update and/or replace a worklet for other workflow system usage. Besides, we provide a semantic description for service function which requires bioinformatics ontology to describe them. In our work we choose to store functional and non-functional services descriptions as semantic description on the service disk entity. This allows reasoning with such information, which is not possible otherwise and is a vital component for service selection and optimization.

In general, worklets may be associated with either an atomic task, or a multiple-instance atomic task. Any number of worklets can form the repertoire of an individual task, and any number of tasks in a particular specification can be associated with the worklet service.

A. Worklet Services Selection

The selection of services inside the worklet depends on the categories, functionality, and ratings of previous interactions. In this work, we use categories as the criterion for the discovery process of those comparable services since they very well express the functionality of services. The discovery process consists basically of: querying BioCatalogue, and using its recent launched API [5]. Each query searches for a specific category and retrieves a set of web services that are annotated with this category. In most cases, the search retrieves more than one service, since there are many services annotated with the same category; however, the set of retrieved services may include services that do not provide exactly the required functionality. This instance is due to the fact that some categories are too general to describe a specific functionality. Among those services we select the services that are comparable based on the services ontology.

Besides, we find "myGrid" ontology [18] the most promising ontology to use, but it can be too big and complex for our purpose. Therefore, we use "myGrid" ontology as a reference and build our own domain ontology only covering the data and analysis required. Our bioinformatics ontology is hence a small subset of the myGrid ontology, which should allow easy migration to the full myGrid ontology if we decide to do so. Every service available in our system is registered with the bioinformatics ontology term(s) that describe its function. Searching for service with specific function becomes equivalent to identifying that function term in the ontology and returning all the services registered with it and its descendant terms.

Also, we use the service ontology for describing services capabilities. For this purpose, we considered both OWL-S [17] and WSMO [8]. While both are capable of describing the service properties that our system requires, we select WSMO because its model of describing non-functional properties matches better our requirements. On the other hand, as there exists many non-functional properties that can be used for service description [19], [20], we mainly focus on reliability, performance, and sensitivity. Those non-functional measures are good indicator of the QoS in our case. Those measures are defined briefly in the following discussion:

- **Reliability R:** measures the availability and stability of Web services. In the bioinformatics context, this can be quantified as the percentage of the up time of the data or analysis services.

$$Reliability = \frac{T_{up}}{T_{total}} \times 100 \tag{1}$$

  Where T_{total} is the total number of attempts trying to use the service, and T_{up} is the number of times the service can be successfully invoked.

- **Performance:** measures the time a Web Service takes to complete a specific task. Many of the bioinformatics analyses are computationally expensive, such as BLASTing against a large data collection and multiple sequence alignment. Slow response is a common experience (sometime up to hours) when a service is requested by a large number of users. Average performance over a long period can be used as an indicators of the service capability. The value of performance is assigned semi-automatically: the system keeps a record of the completion time and input settings of every service execution, the evaluation of service performance P is given by the normalized utility function:

$$P = 0.5 \quad rt \tag{2}$$

  Where rt is the relation between the service response time and the size of the input file, and is given by the following relation

$$rt = \frac{response \quad time}{input \quad size} \tag{3}$$

- **Sensitivity:** refers to the ability to identify all significant information that is related to the input data independent of its quality. Sensitivity is calculated by corroborating results of different services: this repeats the request sent to one service to a different service, and compares the matches given by both. The premise is that services that give top hits are similar enough to validate each other results and, thus, the comparison of different results can identify both true positives and false positives between the best hits.

After we retrieve the set of services that match a specific category and functionality, semantic description is added to those services. The domain ontology provides a biological description of the services, while the service ontology provides the property information among others. Then, we use the service selection algorithm proposed in [21]. The service selection algorithm presented in [21] can be used as tool to help in the selection of Web services based on the available providers and user requirements. In brief the algorithm proceeds as follows, after the selection of the available service providers that serve the user request; the algorithm selects the best set of services as required by the user. The authors rank those set of candidate services and only present those services that most likely solve the user request. A key feature of that approach is that, instead of asking the user for the non-functional properties, the algorithm uses the importance level for QoS parameters from his profile, which makes this algorithm easy to use even for someone not very familiar with the different QoS attributes.

Then, the system allows the user to rate any of the matched services, indicating how relevant or appropriate they are for his request. Besides, a key advantage of the work in [21] is the storing of meta-data about earlier successful services invocations even by other users. Finally, the algorithm keeps services updated by checking their status periodically and providing the user with a report about services usages. In this work, we employ the same algorithm to select the best service based on QoS values.

A workflow specification often contains the physical locations of the resources used. Resources can be (temporarily) unavailable, can be replaced, or can be moved to a different location. This complicates workflow reuse. Delaying the choice of resources until instantiation time would be a solution [22], and is known as late binding [22]. Late binding is supported in our workflow execution since we select the appropriate Web services that can execute the task dynamically at run time instead of hard coding the services that can execute the task during the design of the workflow as described below.

B. Service Selection Movie

In this paper we propose to use a movie set to represent our idea of building the worklet and its selection. The Movie metaphor is introduced in the Dutch Driving Simulator [23] to create an easy-to-understand architecture that supports dynamic generation of traffic scenarios.

In the Movie Metaphor, the world is seen as a movie set in which actors play roles and executes tasks conforming to that role. In the following discussion we demonstrate how this Movie Metaphor approach (see Table I) can be applied to create an easy-to-understand architecture for a workflow task execution. In the movie set, the worklets are the actors and the workflow engine is the director. The role describes the function of the actor required to execute the task [24]. The director controls the set; he selects the tasks to be executed based on the script. The director does not work in isolation but gets help from the casting director to select the actors to execute tasks (known as actor assignment [25]). Like the script of a theatrical play, a workflow specification can be performed (instantiated) more than once and each performance can have different actors involved. Table 1 presents an overview of the terms we borrowed from the movie set and what they stand for and how they can be interpreted in a workflow context.

The actors, director, and casting director interact. Based on the script, the director determines the tasks to be executed. For each of those tasks, the director asks the casting director for an actor to execute it. The casting director searches for an actor in so-called actor repositories, which acts as casting agencies. Multiple actors can be available to play the same role. Based on the task and the role attached to it, the casting director selects a capable actor and delivers him to the director. In workflow terms (see Figure 2), a suitable worklet is selected based on the role (category of worklet) and the task (operation type). The director delegates the task to the selected actor, which then executes the task. This results in 3 possible scenarios:

1) Workflow designers can define a preferred actor for a task. If the casting director finds this preferred actor, it returns this actor.
2) If no preferred actor is set or available and only a single actor is suitable, the casting director selects this actor.
3) If no actors are available, this means that this task has not been considered before. Hence, a new worklet is to be created, and then we create a new actor that represents the new added worklet and store it inside the actor repositories.

Table I: Terms in the movie set and their usage in the workflow paradigm.

Term	Movie Set	In Workflow Context
Script	The story describing the movie	A (hierarchical) workflow specification. A workflow that specifies the tasks to be executed.
Scene	A unit of action, taken at a single location.	A (sub)workflow, consisting of tasks constituting a higher level task which we refer to as a Worklet in our workflow.
Actor	A person with the capabilities to play a certain role.	A resource, i.e. a Worklet that carry certain task.
Role	Specification of a character and its tasks.	A specification (category) of actor required to perform a task.
Director	Person who directs the movie.	The workflow engine; it selects and schedules tasks
Casting director	Person responsible for selecting actors based on the role descriptions.	The component that selects actors based on the role attached to the task.

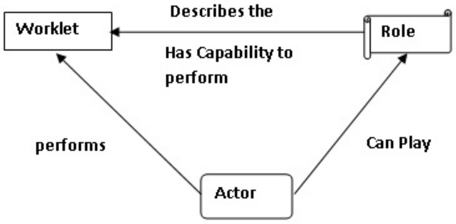


Figure 2: Worklet Role Assignment.

The worklet categories are translated to roles and are registered in our system. Only one role representing that worklet category is registered. To simplify the discovery of roles, an actor repository is created for each worklet (i.e. reference to the worklet).

Worklets are discovered using information about the desired functionality that the user needs. The user specifies the workflow and its script. The casting director then searches for actor(s) (i.e. worklet) using the role attached to the available actors. Note that even with the case that the user specifies the full specification of the workflow he needs, there is still a great deal of abstraction. For example, the algorithm that will be used, the accuracy of the service that provides the result, etc. This alone is a major step forward compared to the current state of in this research direction. In practice, most users will set their own personal preferences for the individual service choices and associated options, and so in most cases the interaction will return higher levels of abstraction in subsequent use. In addition, the worklet service creates a

Table II: Discovered Worklets

Worklet		
Name	Category	ID
Retrieve Protein sequence	Protein Sequence Retrieval	1
Gene prediction	Gene Prediction	2
Protein sequence alignment	Protein sequence alignment	3
Phylogentic tree	Phylogyn	4
Protein sequence analysis	Protein sequence analysis	5
Protein sequence repeats analysis	Protein sequence repeats analysis	6

Table III: Actors for Discovered Worklets and their Roles

Actor				
P	A	N	Role	ID
✓			Retrieves information from the database of protein	1
	✓		produces a list of predicted genes given a sequence of DNA.	2
	✓		Display basic information about a multiple sequence alignment	3
✓			Builds the most accurate phylogenetic tree.	4
✓			A protein sequence and annotation database.	5
		✓		6

process log to provide a complete operational history of each process that is then stored in the system.

V. Use Case

In this section, we develop a bioinformatics case study extracted from [26]. In this way, we can illustrate how to use our approach to guide the user in a real web service discovery tasks. The case study considers biological research that analyzes the presence of specific genes involved in the genesis of Parkinson Disease, called LRRK2 genes, in different organisms. The goal is to know more about the biochemical and cellular functions of these genes. The authors study the presence of the LRRK2 genes in organism "N. Vectensis", since previous studies have shown that this organism is a key organism to trace the origin of these genes. The authors describe the process step-by-step.

We have selected this case study because it describes with details the techniques used in every step, and it could be useful to validate our approach. However, our intention is not to model a concrete case study, but to offer a guide for more general cases. In the rest of this section we present a short description for each step of our approach in order to discover the web services that provide the functionality required by the scientist.

In this case study the user requirement is to build a workflow to obtain a comparison of the LRRK2 genes in different organisms. The steps the user will do manually by himself to build the workflow are [26]:

1) Retrieve the protein sequences of the different domains;
2) Predict the gene structure automatically for the sequences retrieved in Step (1);
3) Align protein sequences to build phylogenetic trees;
4) Build the phylogenetic trees;
5) Analyze the structure of the proteins;

In our model the execution process will begin from the last step in the case of "Analyze the structure of proteins", and continue until no more pre-conditions are needed in the process of "Retrieve protein sequences". The experiments we have made until now suggest that the task descriptions are short and simple sentences not complex.

We consider each of the tasks of the workflow as a scene; for each scene, the casting director searches for an actor who can perform this task. This actor can be: 1) Preferred, 2) Available, or 3) Not available, as shown in Table II, and Table III. For each of the user tasks a worklet is selected. For example, when the user wants to perform Worklet ID=1, the actor that represents the "protein sequence retrieval worklet"' is the matching actor, this actor is set as the *preferred* one, and we select it based on its role.

In case of worklet ID=6 whose actor is not available, in other words there is no worklet that matches user request, in such case a new worklet is created and its category is transformed to the role attached to the actor that will represent this worklet. For each of these worklets, the Web service discovery process is carried out by searching in the BioCatalogue registry services that are annotated with the same category as the user defined tasks; This search retrieves a set of web services per user-defined task, and there is no way to know in advance which one is the most appropriate for the user-defined task. Using service ontology and domain ontology we retrieve a number of comparable services for each worklet as shown in Table IV.

We apply the selection algorithm in [21] on those services and select the best matching services for each task based on a simple user profile values that assume all QoS criteria are of equal importance. Thus, all criteria have the same weighing factor. We then present them to the user to continue with his workflow.

Table IV: Number of Matched Services without Considering QoS

ID	Number of Services
1	73
2	151
3	219
4	21
5	1132
6	122

Table V: Final matched services considering QoS

ID	Number of Services
1	2
2	5
3	4
4	1
5	6
6	4

Consequently, the proposed conceptual abstraction insulates the user from the difficulty of data format conversions and tool compatibility. However, this abstraction does not compromise power, and therefore provides both the novice and bioinformatics expert with an appropriate tool with significant advantages compared to existing alternatives. In addition, the user is certain that the provided services are working correctly during run time. Table V presents the final number of matching services for each of the selected worklets.

VI. Conclusions and Future Work

In this work we provide new interface for workflow execution that simplify the user interaction with workflow systems. The success, of the proposed approach depends on the degree the community standardizes worklet interfaces. For future work, there are many ways in which this research could be directly extended; results might be improved if the user is given the option to perform more advanced searches by explicitly specifying keywords that represent the service or type descriptions. Also, we would like to perform user studies to assess how well the intent-declaration mechanisms work, and to determine if it increases user productivity, which is the ultimate goal. It is also possible that rigorous evaluations would help to identify which of the proposed features should be re-examined. If positive results are achieved, the logical next step would be to incorporate the idea described here into the Taverna workflow client, as this software is used by a relatively large number of scientists to perform actual life-sciences research. Also, we need to generalize the repository used to incorporate other bioinformatics registries.

References

[1] N. R. BP, "Basics for the construction of phylogenetic trees," *Webmed Central BIOLOGY*, p. 12, 2011.

[2] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows." *Bioinformatics*, vol. 20, no. 17, pp. 45–54, 2004.

[3] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, "Kepler: An extensible system for design and execution of scientific workflows," in *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*. Washington, DC, USA: IEEE Computer Society, 2004.

[4] S. Majithia, M. Shields, I. Taylor, and I. Wang, "Triana: A graphical web service composition and execution toolkit," in *Proceedings of the IEEE International Conference on Web Services*, ser. ICWS '04. Washington, DC, USA: IEEE Computer Society, 2004.

[5] C. A. Goble, K. Belhajjame, F. Tanoh, J. Bhagat, K. Wolstencroft, R. Stevens, E. Nzuobontane, H. McWilliam, T. Laurent, and R. Lopez, "Biocatalogue: A curated web service registry for the life science community," *Nature Precedings*, pp. 2–3, 2009.

[6] J. Cardoso, "Quality of service for workflows and web service processes," *Web Semantics Science Services and Agents on the World Wide Web*, vol. 1, no. 3, pp. 281–308, 2004.

[7] K. K. Verdi, H. J. Ellis, and M. R. Gryk, "Conceptual-level workflow modeling of scientific experiments using nmr as a case study." *BMC Bioinformatics*, vol. 8, p. 31, 2007.

[8] D. Roman, U. Keller, H. Lausen, J. de Bruijn, R. Lara, M. Stollberg, A. Polleres, C. Feier, C. Bussler, and D. Fensel, "Web service modeling ontology," *Appl. Ontol.*, vol. 1, pp. 77–106, January 2005.

[9] M. DiBernardo, R. Pottinger, and M. Wilkinson, "Semi-automatic web service composition for the life sciences using the biomoby semantic web framework," *J. of Biomedical Informatics*, vol. 41, pp. 837–847, October 2008.

[10] L. Zhang, Y. Wang, P. Xuan, A. Duvall, J. Lowe, Y. Wang, A. Subramanian, P. Srimani, F. Luo, and Y. Duan, "Sesame: A new bioinformatics semantic workflow design system," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, Dec 2013, pp. 504–508.

[11] Y. Gil, V. Ratnakar, E. Deelman, G. Mehta, and J. Kim, "Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows," in *Proceedings of the 19th National Conference on Innovative Applications of Artificial Intelligence - Volume 2*, ser. IAAI'07, 2007, pp. 1767–1774.

[12] E. Pignotti, P. Edwards, A. Preece, N. Gotts, and G. Polhill, "Enhancing workflow with a semantic description of scientific intent," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds. Springer Berlin Heidelberg, 2008, vol. 5021, pp. 644–658.

[13] U. Srivastava, K. Munagala, J. Widom, and R. Motwani, "Query optimization over web services," in *Proceedings of the 32nd international conference on Very large data bases*, ser. VLDB '06. VLDB Endowment, 2006, pp. 355–366.

[14] G. Canfora, M. Di Penta, R. Esposito, and M. L. Villani, "An approach for qos-aware service composition based on genetic algorithms," in *Proceedings of the 2005 conference on Genetic and evolutionary computation*, ser. GECCO '05. New York, NY, USA: ACM, 2005, pp. 1069–1075.

[15] M. J. Adams, A. H. ter Hofstede, D. Edmond, and W. M. van der Aalst, "Worklets: A service-oriented implementation of dynamic flexibility in workflows," in *the 14th International Conference on Cooperative Information Systems (CoopIS'06)*, R. Meersman and Z. Tari, Eds. Montpellier, France: Springer, 2006, pp. 291–308.

[16] N. Hashmi, S. Lee, and M. P. Cummings, "Abstracting workflows: Unifying bioinformatics task conceptualization and specification through semantic web services," *W3C Workshop on Semantic Web for Life Sciences*, 2004.

[17] P. Romano, D. Marra, and L. Milanesi, "Web services and workflow management for biological resources." *BMC Bioinformatics*, vol. 6, p. 24, 2005.

[18] K. Wolstencroft, P. Alper, D. Hull, C. Wroe, P. W. Lord, R. D. Stevens, and C. A. Goble, "The mygrid ontology: bioinformatics service discovery," *Int. J. Bioinformatics Res. Appl.*, vol. 3, pp. 303–325, September 2007.

[19] M. R. Rodrigues and M. Luck, "Evaluating dynamic services in bioinformatics," *Cooperative Information Agents X*, vol. 4149, pp. 183–197, September 2006, lecture Notes in Artificial Intelligence.

[20] K. Xu, Q. Yu, Q. Liu, J. Zhang, and A. Bouguettaya, "Web service management system for bioinformatics research: a case study," *Service Oriented Computing and Applications*, vol. 5, pp. 1–15, 2011.

[21] W. Nagy, H. M. Mokhtar, and A. El-Bastawissy, "A flexible tool for web service selection," in *Proceedings of the Fifth International Conference on Intelligent Computing and Information Systems*, Ain Shames Univ, Cairo, Egypt, 2011.

[22] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the challenges of scientific workflows," *Computer*, vol. 40, pp. 24–32, December 2007.

[23] I. H. C. Wassink, E. M. A. G. van Dijk, J. Zwiers, A. Nijholt, J. Kuipers, and A. O. Brugman, "Bringing hollywood to the driving school: Dynamic scenario generation in simulations and games," in *INTETAIN*, 2005, pp. 288–292.

[24] I. Wassink, H. Rauwerda, P. E. van der Vet, T. M. Breit, and A. Nijholt, "E-bioflow: Different perspectives on scientific workflows," in *Bioinformatics Research and Development, BIRD 2008, Vienna, Austria*, ser. Communications in Computer and Information Science, M. Elloumi, J. Küng, M. Linial, R. F. Murphy, K. Schneider, and C. Toma, Eds., vol. 13. Berlin: Springer Verlag, 2008, p. 15.

[25] P. Barthelmess and J. Wainer, "Workflow systems: a few definitions and a few suggestions," in *Proceedings of conference on Organizational computing systems*, ser. COCS '95. New York, NY, USA: ACM, 1995, pp. 138–147.

[26] I. Marin, "Ancient origin of the parkinson disease gene," *Journal of Molecular Evolution*, vol. 67, pp. 41–50, 2008.

Timed-Release Hierarchical Identity-Based Encryption

Toru Oshikiri
Graduate School of Engineering
Tokyo Denki University
Tokyo, Japan

Taiichi Saito
Tokyo Denki University
Tokyo, Japan

***Abstract*—We propose a notion of hierarchical identity-based encryption (HIBE) scheme with timed-release encryption (TRE) mechanism, timed-release hierarchical identity-based encryption (TRHIBE), and define its security models. We also show a generic construction of TRHIBE from HIBE and one-time signature, and discuss the security of the constructed scheme.**

***Keywords*—timed-release encryption, hierarchical identity-based encryption, one-time signature**

I. INTRODUCTION

Timed-release encryption (TRE) [1] [2] [3] [4] [5] is an encryption mechanism that allows a receiver to decrypt a ciphertext only after the time that a sender designates.

Timed-release identity-based encryption (TRIBE) [6] is an extension of TRE having a function of identity-based encryption (IBE). In TRIBE, even a legitimate receiver cannot decrypt a ciphertext using secret key until the time designated by the sender. A TRIBE system consists of a key generation center (KGC), a time server (TS), senders and receivers. A sender encrypts a message using an identity of a receiver and a time after which the ciphertext could be decrypted. The KGC generates a secret key corresponding to an identity of a receiver. The TS periodically broadcasts a time signal corresponding to the current time. The receiver decrypts the ciphertext using the secret key and the time signal corresponding to the time designated by the sender. TRIBE systems use identity of user as his/her public key. TRIBE has an advantage that it does not require linking public keys to identity such as PKI.

Timed-release hierarchical identity-based encryption (TRHIBE) is another extension of TRE having a function of hierarchical identity-based encryption (HIBE). In TRHIBE, even a legitimate receiver cannot decrypt a ciphertext using secret key until a time designated by a sender. A TRHIBE system consists of senders, multiple KGCs, a single TS, and receivers. The KGCs and users have a hierarchical structure in which each KGC generates a secret key corresponding to an identity of a child KGC or a child user. Therefore, the load of derivation of users secret keys can be distributed to multiple KGCs. A sender encrypts a message using an identity of a receiver and a time. The TS periodically broadcasts a time signal corresponding to the current time. The receiver decrypts the ciphertext using the secret key and the time signal corresponding to the time designated by the sender.

II. RELATED WORKS

In TRIBE, a user can decrypt a ciphertext only when the user has the receiver's secret key and the time signal generated by TS. Then, if the receiver does not have the time signal or the TS does not have the secret key, they cannot decrypt the ciphertext. In [6], two security models of TRIBE are defined. One is security against malicious receiver, $\mathsf{IND\text{-}ID\text{-}CCA_{CR}}$ security. The other is security against malicious TS, $\mathsf{IND\text{-}ID\text{-}CCA_{TS}}$ security. A generic construction of TRIBE that achieves the security is also shown in [6]. It is a combination of two IBE schemes and a one-time signature scheme, based on "Parallel Encryption" by Dodis-Katz [7], and the security is proved in the standard model.

III. CONTRIBUTION

In this paper, we introduce timed-release hierarchical identity-based encryption (TRHIBE) and define two security models. One is security against malicious receiver, $\mathsf{IND\text{-}hID\text{-}CCA_{CR}}$ security. The other is security against malicious TS, $\mathsf{IND\text{-}hID\text{-}CCA_{TS}}$ security. We also present a generic construction of TRHIBE. It is a combination of two HIBE schemes and a one-time signature scheme, also based on "Parallel Encryption". We see that if the primitive HIBE schemes are $\mathsf{IND\text{-}hID\text{-}CCA}$ secure and the primitive one-time signature scheme is $\mathsf{OT\text{-}sEUF\text{-}CMA}$ secure, then the constructed TRHIBE scheme is $\mathsf{IND\text{-}hID\text{-}CCA_{CR}}$ secure and $\mathsf{IND\text{-}hID\text{-}CCA_{TS}}$ secure in the standard model.

IV. PRELIMINARIES

In this section, we review hierarchical identity-based encryption (HIBE) and one-time signature, which we use later.

A. Hierarchical Identity-Based Encryption

In an HIBE scheme, the single KGC functionality of generating secret keys is divided into partial ones and they are delegated to multiple KGCs. If a KGC is assigned an identity vector, $\mathsf{ID}^{(k-1)} = (I_1, ..., I_{k-1})$, and given a secret key, $d_{\mathsf{ID}^{(k-1)}}$, corresponding to the identity vector, then it can generate a secret key, $d_{\mathsf{ID}^{(k)}}$, corresponding to an identity vector, $\mathsf{ID}^{(k)} = (I_1, ..., I_{k-1}, I_k)$. We may denote an identity by ID if we need not to specify its hierarchy depth.

Let λ be a security parameter and ℓ be a maximum depth of hierarchy. An *hierarchical identity-based encryption scheme* $\mathcal{HIBE}$ consists of five probabilistic polynomial-time algorithms $\mathcal{HIBE}$ =(HIBE.Setup, HIBE.Ext, HIBE.Del, HIBE.Enc, HIBE.Dec). The setup algorithm HIBE.Setup takes λ and ℓ as input, and outputs a public parameter $params$ and a master secret key msk. The extract algorithm HIBE.Ext takes $params$, msk, and an identity $\mathsf{ID}^{(k)} = (\mathsf{I}_1, \ldots, \mathsf{I}_k)$ as inputs, and outputs a decryption key $d_{\mathsf{ID}^{(k)}}$. The delegate algorithm HIBE.Del takes $params, \mathsf{ID}^{(k)}, d_{\mathsf{ID}^{(k)}}$ and an identity $\mathsf{ID}^{(k+1)}$ as inputs, and outputs a decryption key $d_{\mathsf{ID}^{(k+1)}}$. The encryption algorithm HIBE.Enc takes $params, \mathsf{ID}$, a message m as inputs, and outputs a ciphertext c. The decryption algorithm HIBE.Dec takes $params$, a ciphertext c and a decryption key d_{ID} as inputs, and outputs the plaintext m' or $\perp$. These algorithms are assumed to satisfy that if $(params, msk) = \mathsf{HIBE.Setup}(\lambda)$ and $d_{\mathsf{ID}} = \mathsf{HIBE.Ext}(params, msk, \mathsf{ID})$ or $d_{\mathsf{ID}^{(k)}} = \mathsf{HIBE.Del}(params, \mathsf{ID}^{(k-1)}, d_{\mathsf{ID}^{(k-1)}},\ \mathsf{ID}^{(k)})$ for $k \leq n$, then $\mathsf{HIBE.Dec}(params, d_{\mathsf{ID}}, \mathsf{HIBE.Enc}(params, \mathsf{ID}, m)) = m$ for any m.

1) IND-hID-CCA *Security:* We review a standard security notion for HIBE: indistinguishability against adaptive hierarchical identity and chosen ciphertext attacks (IND-hID-CCA) security [8] [9]. We here describe the IND-hID-CCA security for HIBE scheme $\mathcal{HIBE}$ based on the following IND-hID-CCA game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs $(params, msk) \leftarrow \mathsf{HIBE.Setup}(\lambda, \ell)$. $\mathcal{C}$ sends $params$ to $\mathcal{A}$ and keeps msk secret.

Phase1

$\mathcal{A}$ can adaptively issue *extraction queries* ID and *decryption queries* (ID, c). $\mathcal{C}$ responds to an extraction query ID by running $d_{\mathsf{ID}_j} = \mathsf{HIBE.Ext}(params, msk, \mathsf{ID})$ and returning d_{ID} to $\mathcal{A}$. $\mathcal{C}$ responds to a decryption query (ID, c) by running $d_{\mathsf{ID}} = \mathsf{HIBE.Ext}(params, msk, \mathsf{ID})$ and $m' = \mathsf{HIBE.Dec}(d_{\mathsf{ID}}, c)$, and returning m' to $\mathcal{A}$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, and an identity to be challenged ID^* to $\mathcal{C}$. The challenge identity ID^* must differ from any ID issued as extraction query in **Phase1**, and any its prefixes. $\mathcal{C}$ randomly chooses $b \in \{0, 1\}$ and sends a challenge ciphertext $c^* = \mathsf{HIBE.Enc}(params, \mathsf{ID}^*, m_b)$ to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue extraction queries ID and decryption queries (ID, c) in the same way as in **Phase1** except that the extraction queries ID must differ from the challenge identity ID^* and its prefixes, and decryption queries (ID, c) must differ from the pair (ID^*, c^*).

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the IND-hID-CCA game as $Adv^{\mathsf{IND\text{-}hID\text{-}CCA}}_{\mathcal{HIBE},\mathcal{A}}(\lambda) = |2\Pr[b = b'] - 1|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the HIBE scheme $\mathcal{HIBE}$ is IND-hID-CCA *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{IND\text{-}hID\text{-}CCA}}_{\mathcal{HIBE},\mathcal{A}}(\lambda)$ is negligible in λ.

B. Signature

Let λ be a security parameter. $\mathcal{A}$ *signature* scheme $\mathcal{SIG}$ consists of three probabilistic polynomial-time algorithms $\mathcal{SIG} = (\mathsf{SigGen}, \mathsf{Sign}, \mathsf{Verify})$. The key generation algorithm SigGen takes λ as input, and outputs a signing key sk and a verification key vk. The signing algorithm Sign takes sk and a message m as inputs , and outputs a signature σ. The verification algorithm Verify takes vk, a message m, and a signature σ as inputs, and outputs `accept` or `reject`. These algorithms are assumed to satisfy that if $(sk, vk) = \mathsf{SigGen}(\lambda)$ then $\mathsf{Verify}(vk, m, \mathsf{Sign}(sk, m)) = \mathtt{accept}$ for any m.

1) OT-sEUF-CMA *Security:* We review a security notion for one-time signature scheme: one-time strong existential unforgeability against chosen message attacks (OT-sEUF-CMA) security [10]. We here describe the OT-sEUF-CMA security for signature scheme $\mathcal{SIG}$ based on the following OT-sEUF-CMA game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs the (*sk*, *vk*) $\leftarrow$ SigGen(λ). $\mathcal{C}$ sends vk to $\mathcal{A}$ and keeps sk secret.

Query

$\mathcal{A}$ can issue a signing query m to $\mathcal{C}$ only once. $\mathcal{C}$ responds to the singing query m by running $\sigma = \mathsf{Sign}(vk, m)$ and returning σ to $\mathcal{A}$.

Forge

$\mathcal{A}$ outputs a pair (m^*, σ^*).

We define the advantage of $\mathcal{A}$ in the OT-sEUF-CMA game as $Adv^{\mathsf{OT\text{-}sEUF\text{-}CMA}}_{\mathcal{SIG},\mathcal{A}}(\lambda)$ = $\Pr[\mathsf{Verify}(vk,\ m^*,\ \sigma^*) = \mathtt{accept} \land (m, \sigma) \neq (m^*, \sigma^*)]$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the signature scheme $\mathcal{SIG}$ is OT-sEUF-CMA *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{OT\text{-}sEUF\text{-}CMA}}_{\mathcal{SIG},\mathcal{A}}(\lambda)$ is negligible in λ.

V. Timed-Release Hierarchical Identity-Based Encryption(TRHIBE)

In this section, we introduce timed-release hierarchical identity-based encryption(TRHIBE) scheme and define its security models.

A TRHIBE system consists of a single TS, multiple KGCs and multiple users connected through a communication network. The time server periodically broadcasts a time signal corresponding to the current time, and all users can receive the time signal. The single KGC functionality of generating secret keys is divided into partial ones and they are delegated to multiple KGCs. If a KGC is assigned an identity vector, $\mathsf{ID}^{(k-1)} = (I_1, ..., I_{k-1})$, and given a secret key, $d_{\mathsf{ID}^{(k-1)}}$, corresponding to the identity vector, then it can generate a secret key, $d_{\mathsf{ID}^{(k)}}$, corresponding to an identity vector, $\mathsf{ID}^{(k)} = (I_1, ..., I_{k-1}, I_k)$. We may denote an identity by ID if we need not to specify its hierarchy depth. A user (sender) encrypts a plaintext, designating another user (receiver) who can decrypt the ciphertext and a time only after which the ciphertext can be decrypted. The receiver can decrypt the ciphertext with the secret key that he/she has and the time signal that the time server broadcasts at the designated time.

Let λ be a security parameter and ℓ be a maximum depth of system. An *timed-release hierarchical identity-based encryption scheme* $\mathcal{TRHIBE}$ consists of seven probabilistic polynomial-time algorithms $\mathcal{TRHIBE}$=(TS_Setup, KGC_Setup, Release, Extract, Delegate, Encrypt, Decrypt). The time server's setup algorithm TS_Setup takes λ as input, and outputs a public key tpk and the corresponding secret key tsk. The key generation center's setup algorithm KGC_Setup takes λ and the depth ℓ as input, and outputs a public parameter $params$ and a master secret key msk. The release algorithm Release takes tpk, tsk and a time period T as inputs, and outputs a time signal d_T. The extract algorithm Extract takes $params$, msk, and an identity $\mathsf{ID}^{(k)} = (\mathsf{I}_1, \ldots, \mathsf{I}_k)$ as inputs, and outputs a decryption key $d_{\mathsf{ID}^{(k)}}$. The delegate algorithm Delegate takes $params$, $\mathsf{ID}^{(k)}$, $d_{\mathsf{ID}^{(k)}}$ and an identity $\mathsf{ID}^{(k+1)}$ as inputs, and outputs a decryption key $d_{\mathsf{ID}^{(k+1)}}$. The encryption algorithm Encrypt takes $tpk, params$, T, and ID, and a message m as inputs, and outputs a ciphertext c. The decryption algorithm Decrypt takes as inputs $tpk, params$, a ciphertext c', d_T, a user's secret key d_{ID}, and outputs the plaintext m' or $\perp$. These algorithms are assumed to satisfy that Decrypt(tpk, $params$, d_T, d_{ID}, Encrypt(tpk, $params$, T, ID, m)) = m holds for any m, if (tpk, tsk) = TS_Setup(λ), ($params$, msk) = KGC_Setup(λ, ℓ), s_T = TR.Release(tpk, tsk, T), and d_{ID} = HIBE.Ext($params$, msk, ID) hold, and that $d_{\mathsf{ID}^{(n)}}$ = HIBE.Ext($params$, msk, $\mathsf{ID}^{(n)}$) and $d_{\mathsf{ID}^{(k)}}$ = HIBE.Del($params$, $\mathsf{ID}^{(k-1)}$, $d_{\mathsf{ID}^{(k-1)}}$, $\mathsf{ID}^{(k)}$) for $k \leq n$ hold.

A. Security

We can consider security against malicious TS and security against malicious receiver.

1) IND-hID-CCA$_{\mathsf{TS}}$ *Security.:* We introduce a security notion for TRHIBE: *indistinguishability against adaptive hierarchical identity and chosen ciphertext attacks by time-servers (*IND-hID-CCA$_{\mathsf{TS}}$*) security*. This security ensures that a malicious time server, who has a secret key tsk, cannot obtain any information of message from ciphertext without decryption key d_{ID}. We here describe the IND-hID-CCA$_{\mathsf{TS}}$ security for a TRHIBE scheme $\mathcal{TRHIBE}$ based on the following IND-hID-CCA$_{\mathsf{TS}}$ game between a challenger $\mathcal{C}$ and adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs (tpk, tsk) $\leftarrow$ TS_Setup(λ) and ($params$, msk) $\leftarrow$ KGC_Setup(λ, ℓ). $\mathcal{C}$ sends tpk, tsk and $params$ to $\mathcal{A}$ and keeps msk secret.

Phase1

$\mathcal{A}$ can adaptively issue extraction queries ID and decryption queries (T, ID, c). $\mathcal{C}$ responds to an extraction query ID by running d_{ID} = Extract($params$, msk, ID) and returning d_{ID} to $\mathcal{A}$. $\mathcal{C}$ responds to a decryption query (T, ID, c) by running d_T = Release(tpk, tsk, T), d_{ID} = Extract($params$, msk, ID) and c = Decrypt(d_T, d_{ID}, c) , and returning c to $\mathcal{A}$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, a time period T^* and an identity to be challenged ID* to $\mathcal{C}$. The challenge identity ID* must differ from any ID issued as extraction queries in **Phase1** and any its prefixes. $\mathcal{C}$ randomly chooses $b \in \{0, 1\}$ and sends a challenge ciphertext $c^* = \mathsf{Encrypt}(tpk, params, T^*, \mathsf{ID}^*, m_b)$ to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue extraction queries ID and decryption queries (T, ID, c) in the same way as **Phase1** except that the extraction queries ID must differ from the challenge identity ID* and its prefixes, and the decryption queries (T, ID, c) must differ from the tuple $(T^*, \mathsf{ID}^*, c^*)$.

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the IND-hID-CCA$_{\mathsf{TS}}$ game as $Adv^{\mathsf{IND\text{-}hID\text{-}CCA_{TS}}}_{\mathcal{TRHIBE},\mathcal{A}}(\lambda) = |2\Pr[b = b'] - 1|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the TRIBE scheme $\mathcal{TRHIBE}$ is IND-hID-CCA$_{\mathsf{TS}}$ *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\mathsf{IND\text{-}hID\text{-}CCA_{TS}}}_{\mathcal{TRHIBE},\mathcal{A}}(\lambda)$ is negligible in λ.

2) IND-hID-CCA$_{\mathsf{CR}}$ *Security.:* We introduce another security notion for TRIBE: *indistinguishability against adaptive hierarchical identity and chosen ciphertext attacks by curious receiver (*IND-hID-CCA$_{\mathsf{CR}}$*) security*. This security ensures that a receiver who has a decryption key d_{ID} cannot obtain any information of message from ciphertext without time signal d_T. We here describe the IND-hID-CCA$_{\mathsf{CR}}$ security for a TRIBE scheme $\mathcal{TRHIBE}$ based on the following IND-hID-CCA$_{\mathsf{CR}}$ game between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$.

Setup

$\mathcal{C}$ runs (tpk, tsk) $\leftarrow$ TS_Setup(λ) and ($params$, msk) $\leftarrow$ KGC_Setup(λ, ℓ). $\mathcal{C}$ sends $params$, msk and tpk to $\mathcal{A}$ and keeps tsk secret.

Phase1

$\mathcal{A}$ can adaptively issue release queries T and decryption queries (T, ID, c). $\mathcal{C}$ responds to a release query T by running d_T = Release(tpk, tsk, T) and returning d_T to $\mathcal{A}$. $\mathcal{C}$ responds to a decryption query (T, ID, c) by running d_T = Release(tpk, tsk, T), $d_{\mathsf{ID}} = \mathsf{Extract}(paramsmsk, \mathsf{ID})$ and $c = \mathsf{Decrypt}(d_T, d_{\mathsf{ID}}, c)$, and returning c to $\mathcal{A}$.

Challenge

$\mathcal{A}$ sends two messages m_0, m_1 such that $|m_0| = |m_1|$, a time period T^* and an identity ID* to be challenged to $\mathcal{C}$. The challenge time period T^* must differ from any T issued as release queries in Phase1. $\mathcal{C}$ randomly chooses $b \in \{0, 1\}$ and sends a challenge ciphertext c^* = Encrypt(tpk, $params$, T^*, ID*, m_b) to $\mathcal{A}$.

Phase2

$\mathcal{A}$ can adaptively issue release queries T and decryption queries (T, ID, c) in the same way as **Phase1** except that the release query T must differ from the challenge time period T^*, and the decryption queries (T, ID, c) must differ from the tuple $(T^*, \mathsf{ID}^*, c^*)$.

Guess

$\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$ and wins if $b = b'$.

We define an advantage of $\mathcal{A}$ in the IND-hID-CCA$_{\text{CR}}$ game as $Adv^{\text{IND-hID-CCA}_{\text{CR}}}_{\mathcal{TRHIBE},\mathcal{A}}(\lambda) = |2\Pr[b = b'] - 1|$, in which the probability is taken over the random coins used by $\mathcal{C}$ and $\mathcal{A}$. We say that the TRIBE scheme $\mathcal{TRHIBE}$ is IND-hID-CCA$_{\text{CR}}$ *secure* if, for any probabilistic polynomial-time adversary $\mathcal{A}$, the function $Adv^{\text{IND-hID-CCA}_{\text{CR}}}_{\mathcal{TRHIBE},\mathcal{A}}(\lambda)$ is negligible in λ.

VI. Construction of TRHIBE

Here we present a generic construction of TRHIBE scheme from two HIBE schemes, and a one-time signature scheme.

A. Construction

Let Π = (HIBE.Setup, HIBE.Ext, HIBE.Del, HIBE.Enc, HIBE.Dec) and Π' = (HIBE$'$.Setup, HIBE$'$.Ext, HIBE$'$.Del, HIBE$'$.Enc, HIBE$'$.Dec) be hierarchical identity-based encryption schemes, and Σ = (SigGen, Sign, Verify) be a one-time signature scheme.

A TRHIBE scheme Γ = (TS_Setup, KGC_Setup, Release, Extract, Encrypt, Decrypt) is constructed as follows.

Time server setup TS_Setup(λ)**:**
Step 1: Run HIBE.Setup(λ, 1) to generate $(params, msk)$.
Step 2: Set $tpk = params$ and $tsk = msk$.
Step 3: Return (tpk, tsk).

Key generation center setup KGC_Setup(λ, ℓ)**:**
Step 1: Run HIBE$'$.Setup(λ, ℓ) to generate $(params, msk)$.
Step 2: Return $(params, msk)$.

Release Release(tpk, tsk, T)**:**
Step 1: Run HIBE.Ext(tpk, tsk, T) to obtain d_T.
Step 2: Return d_T.

Extraction Extract($params, msk$, ID)**:**
Step 1: Run HIBE$'$.Ext($params, msk$, ID) to obtain d_{ID_j}.
Step 2: Return d_{ID_j}.

Delegate($params, \mathsf{ID}^{(k)}, d_{\mathsf{ID}^{(k)}}, \mathsf{ID}^{(k+1)}$)**:**
Step 1: Run HIBE$'$.Del($params, \mathsf{ID}^{(k)}, d_{\mathsf{ID}^{(k)}}, \mathsf{ID}^{(k+1)}$) to obtain $d_{\mathsf{ID}^{(k+1)}}$.
Step 2: Return $d_{\mathsf{ID}^{(k+1)}}$.

Encryption Encrypt($tpk, params, m, T$, ID)**:**
Step 1: Run SigGen(λ) to generate (sk, vk).
Step 2: Randomly choose $s_1 \in \{0,1\}^{|m|}$.
Step 3: Compute $s_2 = m \oplus s_1$.
Step 4: Compute c_1 = HIBE.Enc($tpk, s_1||vk, T$).
Step 5: Compute c_2 = HIBE$'$.Enc($params, s_2||vk$, ID).
Step 6: Compute σ = Sign($sk, c_1||c_2||T||$ID).
Step 7: Set $c = (c_1, c_2, T, \mathsf{ID}, vk, \sigma)$.
Step 8: Return c.

Decryption Decrypt($tpk, params, c, d_T, d_{\mathsf{ID}}$)**:**
Step 1: Parse c as $c = (c_1, c_2, T, \mathsf{ID}, vk, \sigma)$.
Step 2: If Verify($vk, c_1||c_2||T||$ID$, \sigma$)= `reject` then return $\perp$ and stop.
Step 3: Compute $s_1||vk'$ = HIBE.Dec(tpk, c_1, d_T).
Step 4: Compute $s_2||vk''$ = HIBE$'$.Dec($params, c_2, d_{\mathsf{ID}}$).
Step 5: If $vk = vk' = vk''$ then return $m = s_1 \oplus s_2$, else return $\perp$.

B. Security of TRHIBE.

1) IND-hID-CCA$_{\text{TS}}$ secure:

Theorem 1: If Π' is an IND-hID-CCA secure hierarchical identity-based encryption scheme and Σ is a OT-sEUF-CMA secure one-time signature scheme, then Γ is an IND-hID-CCA$_{\text{TS}}$ secure timed-release hierarchical identity-based encryption scheme.

Proof(Theorem 1) Suppose $\mathcal{A}$ is an adversary that breaks the IND-hID-CCA$_{\text{TS}}$ security of Γ. We construct a simulator $\mathcal{B}$ which breaks the IND-hID-CCA security of the HIBE scheme Π' using $\mathcal{A}$. Say a ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk, \sigma)$ is *valid* if Verify($vk, c_1||c_2||T||$ID$, \sigma$) = `accept`. Let $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ be the challenge ciphertext. Let `Forge` denote the event that $\mathcal{A}$ submits a valid ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk^*, \sigma)$ as a decryption query to $\mathcal{C}$ in the **Phase2**, and `Succ` denote the event that $\mathcal{B}$ wins the IND-hID-CCA game. We prove the following claims.

Claim 1: $\Pr[\texttt{Forge}]$ is negligible.

Claim 2: $\Pr[\texttt{Succ}|\overline{\texttt{Forge}}] = Adv^{\text{IND-hID-CCA}_{\text{TS}}}_{\Gamma,\mathcal{A}} + \frac{1}{2}$

Proof(Claim 1) We assume `Forge` occurs. Then, we construct a forger $\mathcal{F}$ who breaks OT-sEUF-CMA security of the one-time signature scheme Σ, from $\mathcal{A}$. The description of $\mathcal{F}$ is as follows.

Setup
$\mathcal{F}$ receives vk^* from $\mathcal{C}$. Then $\mathcal{F}$ runs (tpk, tsk) $\leftarrow$ TS_Setup(λ) and ($params$, msk) $\leftarrow$ KGC_Setup(λ, ℓ). $\mathcal{F}$ sends tpk, tsk and $params$ to $\mathcal{A}$ and keeps msk.

Query
$\mathcal{F}$ can respond to extract queries and decryption queries of $\mathcal{A}$ since $\mathcal{F}$ has tsk and msk. If $\mathcal{A}$ happens to issue a valid ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk^*, \sigma)$ as decryption query to $\mathcal{F}$ before **Challenge** in the IND-hID-CCA$_{\text{TS}}$ game, then $\mathcal{F}$ simply outputs $(c_1||c_2||T||\mathsf{ID}, \sigma)$ as forgery and stops.

Challenge
If $\mathcal{A}$ outputs $(m_0, m_1, T^*, \mathsf{ID}^*)$ as challenge, $\mathcal{F}$ randomly chooses $s_1 \in \{0,1\}^{|m|}$ and $b \in \{0,1\}$, and computes $s_2 = m_b \oplus s_1$. Then $\mathcal{F}$ computes c_1^* = HIBE.Enc($tpk, s_1||vk^*, T^*$) and c_2^* = HIBE'.Enc($params, s_2||vk^*$, ID*), then issues $m^* = (c_1||c_2||T^*||\mathsf{ID}^*)$ as signing query to $\mathcal{C}$ and obtains σ^*. Finally $\mathcal{F}$ returns $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ as the challenge ciphertext to $\mathcal{A}$.

Forge
If $\mathcal{A}$ issues a valid ciphertxt c = (c_1, c_2, T, ID, vk^*, σ) as decryption query, then $\mathcal{F}$ outputs $(c_1||c_2||T||\mathsf{ID}^*, \sigma)$ as forgery.

$\mathcal{F}$ can forge the signature if $\mathcal{A}$ issues a decryption query that causes the event `Forge`. It, however, contradicts that Σ is OT-sEUF-CMA secure. Thus, $\Pr[\texttt{Forge}]$ is negligible. □

Proof(Claim 2) We construct an adversary $\mathcal{B}$ who breaks IND-hID-CCA security of the HIBE scheme Π' using $\mathcal{A}$. The description of $\mathcal{B}$ is as follows.

Setup

$\mathcal{B}$ receives $params$ from $\mathcal{C}$. Then $\mathcal{B}$ runs (*tpk*, *tsk*) $\leftarrow$ $\mathsf{TS\_Setup}(\lambda)$ and sends tpk, tsk and $params$ to $\mathcal{A}$.

Phase1

$\mathcal{B}$ responds to $\mathcal{A}$'s extraction query ID by issuing ID as $\mathcal{B}$'s extraction query to $\mathcal{C}$ and obtaining d_{ID} from $\mathcal{C}$ and returning d_{ID} to $\mathcal{A}$. $\mathcal{B}$ responds to $\mathcal{A}$'s decryption query c as follows. If $\mathsf{Verify}(vk, c_1||c_2||T||\mathsf{ID}, \sigma) = \mathtt{reject}$, then $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$. Otherwise $\mathcal{B}$ runs $s_1||vk' \leftarrow$ $\mathsf{HIBE.Dec}(c_1, d_T)$ and issues decryption query (c_2, ID) to $\mathcal{C}$ and obtains $s_2||vk''$. $\mathcal{B}$ returns $m = s_1 \oplus s_2$ to $\mathcal{A}$ if $vk = vk' = vk''$, and otherwise $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$.

Challenge

If $\mathcal{A}$ outputs (m_0, m_1, T^*, ID^*) as challenge, $\mathcal{B}$ runs (sk^*, vk^*) $\leftarrow$ $\mathsf{SigGen}(\lambda)$ and randomly chooses $s_1 \in \{0,1\}^{|m|}$ and runs $c_1^* =$ $\mathsf{HIBE.Enc}(tpk, r||vk^*, T^*)$. Then $\mathcal{B}$ computes $M_0 = (m_0 \oplus r||vk^*)$ and $M_1 = (m_1 \oplus r||vk^*)$, and issues $(M_0, M_1, \mathsf{ID}^*)$ as $\mathcal{B}$'s challenge to $\mathcal{C}$ and obtains cyphertext c_2^*. $\mathcal{B}$ runs $\sigma^* = \mathsf{Sign}(sk^*, c_1^*||c_2^*||T^*||\mathsf{ID}^*)$ and returns $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ as challenge ciphertext to $\mathcal{A}$.

Phase2

$\mathcal{B}$ responds to $\mathcal{A}$'s extraction query ID in the same way as in **Phase1**. $\mathcal{B}$ responds to $\mathcal{A}$'s decryption query as follows. The followings are done in a sequential way.

Step1

If $\mathsf{Verify}(vk, c_1||c_2||T||\mathsf{ID}||, \sigma) = \mathtt{reject}$, then $\mathcal{B}$ returns $\perp$ and skips **step2**$\sim$**4**.

Step2

If $vk = vk^*$, then $\mathcal{B}$ stops the simulation and outputs a random bit b'.

Step3

If $(c_2, \mathsf{ID}) = (c_2^*, \mathsf{ID}^*)$, then $\mathcal{B}$ returns $\perp$ and skips **step4**.

Step4

$\mathcal{B}$ responds in the same way as in **Phase1**.

Guess If $\mathcal{A}$ outputs a bit, then $\mathcal{B}$ outputs a same bit as its guess.

We examine the $\mathcal{B}$'s simulation of the response to decryption queries in **Phase2**. In the case of $\mathsf{Verify} = \mathtt{reject}$ in **Step1**, $\mathcal{B}$ returns $\perp$ in the same way as in our decryption algorithm, and then it perfectly simulates the challenger in IND-hID-CCA$_{\mathsf{TS}}$ game. In the case of $vk = vk^*$ in **Step2**, the event $\mathtt{Forge}$ occurs. In the case of $(c_2, \mathsf{ID}) = (c_2^*, \mathsf{ID}^*)$ in **Step3**, since c_2 equals to c_2^*, the decryption of c_2 is $M_0 = (m_0 \oplus r||vk^*)$ or $M_1 = (m_1 \oplus r||vk^*)$. However, since $vk \neq vk^*$, the decryption of c is $\perp$, and then $\mathcal{B}$ simulates perfectly. In the case of $(c_2, \mathsf{ID}) \neq (c_2^*, \mathsf{ID}^*)$, $\mathcal{B}$ can issue the valid decryption query (c_2, ID) to $\mathcal{C}$.

If the event $\mathtt{Forge}$ does not occurs, $\mathcal{B}$ perfectly simulates the challengers in the IND-hID-CCA$_{\mathsf{TS}}$ game and wins the IND-hID-CCA game with the same probability that $\mathcal{A}$ wins the IND-hID-CCA$_{\mathsf{TS}}$ game, i.e., $\Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] = Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{TS}}} + \frac{1}{2}$. □

We see that

$$\begin{aligned}
\Pr[\mathtt{Succ}] &\geq \Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot \Pr[\overline{\mathtt{Forge}}] \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot (1 - \Pr[\mathtt{Forge}]) \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] - \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot \Pr[\mathtt{Forge}] \\
&\geq \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] - \Pr[\mathtt{Forge}],
\end{aligned}$$

then, from **Claim 2**, we have that

$$\Pr[\mathtt{Succ}] \geq Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{TS}}} + \frac{1}{2} - \Pr[\mathtt{Forge}].$$

If $Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{TS}}}$ is not negligible, $Adv_{\Pi,\mathcal{B}}^{\mathsf{IND\text{-}hID\text{-}CCA}} = |\Pr[\mathtt{Succ}] - \frac{1}{2}|$ is not negligible from **Claim 1**, and it contradicts our assumption. This completes the proof of **Theorem 1**. □

2) IND-hID-CCA$_{\mathsf{CR}}$ *secure:*

Theorem 2: If Π is an IND-hID-CCA secure hierarchical identity-based encryption scheme and Σ is a OT-sEUF-CMA secure one-time signature scheme, then Γ is an IND-hID-CCA$_{\mathsf{CR}}$ secure timed-release hierarchical identity-based encryption scheme.

Proof(Theorem 2) Suppose $\mathcal{A}$ is an adversary that breaks the IND-hID-CCA$_{\mathsf{TS}}$ security of Γ. We construct a simulator $\mathcal{B}$ which breaks the IND-hID-CCA security of the HIBE scheme Π using $\mathcal{A}$. Say a ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk, \sigma)$ is *valid* if $\mathsf{Verify}(vk, c_1||c_2||T||\mathsf{ID}, \sigma) = \mathtt{accept}$. Let $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ be the challenge ciphertext. Let $\mathtt{Forge}$ denote the event that $\mathcal{A}$ submits a valid ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk^*, \sigma)$ as a decryption query to $\mathcal{C}$ in the **Phase2**, and $\mathtt{Succ}$ denote the event that $\mathcal{B}$ wins the IND-hID-CCA game. We prove the following claims.

Claim 3: $\Pr[\mathtt{Forge}]$ is negligible.

Claim 4: $\Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] = Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{CR}}} + \frac{1}{2}$

Proof(Claim 3) We assume $\mathtt{Forge}$ occurs. Then, We construct a forger $\mathcal{F}$ who breaks OT-sEUF-CMA security of the one-time signature scheme Σ, from $\mathcal{A}$. The description of $\mathcal{F}$ is as follows.

Setup

$\mathcal{F}$ receives vk^* from $\mathcal{C}$. Then $\mathcal{F}$ runs (*tpk*, *tsk*) $\leftarrow$ $\mathsf{TS\_Setup}(\lambda)$ and (*params*, *msk*) $\leftarrow$ $\mathsf{KGC\_Setup}(\lambda, \ell)$. $\mathcal{F}$ sends $params, msk$ and tpk to $\mathcal{A}$ and keeps tsk.

Query

$\mathcal{F}$ can respond to extract queries and decryption queries of $\mathcal{A}$ since $\mathcal{F}$ has tsk and msk. If $\mathcal{A}$ happens to issue a valid ciphertext $c = (c_1, c_2, T, \mathsf{ID}, vk^*, \sigma)$ as decryption query to $\mathcal{F}$ before **Challenge** in the IND-hID-CCA$_{\mathsf{TS}}$

game, then $\mathcal{F}$ simply outputs $(c_1||c_2||T||\mathsf{ID}, \sigma)$ as forgery and stops.

Challenge

If $\mathcal{A}$ outputs $(m_0, m_1, T^*, \mathsf{ID}^*)$ as challenge , $\mathcal{F}$ randomly chooses $s_1 \in \{0,1\}^{|m|}$ and $b \in \{0,1\}$, and computes $s_2 = m_b \oplus s_1$. Then $\mathcal{F}$ computes c_1^* = HIBE.Enc($tpk, s_1||vk^*, T^*$) and c_2^* = HIBE'.Enc($params, s_2||vk^*, \mathsf{ID}^*$), then issues $m^* = (c_1||c_2||T^*||\mathsf{ID}^*)$ as signing query to $\mathcal{C}$ and obtains σ^*. Finally $\mathcal{F}$ returns $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ as the challenge ciphertext to $\mathcal{A}$.

Forge

If $\mathcal{A}$ issues a valid ciphertxt $c = (c_1, c_2, T, \mathsf{ID}, vk^*, \sigma)$ as decryption query, then $\mathcal{F}$ outputs $(c_1||c_2||T||\mathsf{ID}^*, \sigma)$ as forgery.

$\mathcal{F}$ can forge the signature if $\mathcal{A}$ issues a decryption query that causes the event Forge. It, however, contradicts that Σ is OT-sEUF-CMA secure. Thus, $\Pr[\mathtt{Forge}]$ is negligible. □

Proof(Claim 4) We construct an adversary $\mathcal{B}$ who breaks IND-hID-CCA security of the HIBE scheme Π using $\mathcal{A}$. The description of $\mathcal{B}$ is as follows.

Setup

$\mathcal{B}$ receives $params$ from $\mathcal{C}$. We call this $params$ tpk. Then $\mathcal{B}$ runs (*params*, *msk*) $\leftarrow$ KGC_Setup(λ, ℓ) and sends $params, msk$ and tpk to $\mathcal{A}$.

Phase1

$\mathcal{B}$ responds to $\mathcal{A}$'s release query T by issuing T as $\mathcal{B}$'s extraction query to $\mathcal{C}$ and obtaining d_T from $\mathcal{C}$ and returning d_T to $\mathcal{A}$. $\mathcal{B}$ responds to $\mathcal{A}$'s decryption query c as follows. If $\mathsf{Verify}(vk, c_1||c_2||T||\mathsf{ID}, \sigma) = \mathtt{reject}$, then $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$. Otherwise $\mathcal{B}$ runs $s_2||vk' \leftarrow$ HIBE.Dec(c_2, d_{ID}) and issues decryption query (c_1, T) to $\mathcal{C}$ and obtains $s_1||vk''$. $\mathcal{B}$ returns $m = s_1 \oplus s_2$ to $\mathcal{A}$ if $vk = vk' = vk''$, and otherwise $\mathcal{B}$ returns $\perp$ to $\mathcal{A}$.

Challenge

If $\mathcal{A}$ outputs $(m_0, m_1, T^*, \mathsf{ID}^*)$ as challenge, $\mathcal{B}$ runs $(sk^*, vk^*) \leftarrow$ SigGen(λ) and randomly chooses $s_1 \in \{0,1\}^{|m|}$ and runs c_1^* = HIBE.Enc($params, r||vk^*, \mathsf{ID}^*$). Then $\mathcal{B}$ computes $M_0 = (m_0 \oplus r||vk^*)$ and $M_1 = (m_1 \oplus r||vk^*)$, and issues (M_0, M_1, T^*) as $\mathcal{B}$'s challenge to $\mathcal{C}$ and obtains cyphertext c_2^*. $\mathcal{B}$ runs $\sigma^* = \mathsf{Sign}(sk^*, c_1^*||c_2^*||T^*||\mathsf{ID}^*)$ and returns $c^* = (c_1^*, c_2^*, T^*, \mathsf{ID}^*, vk^*, \sigma^*)$ as challenge ciphertext to $\mathcal{A}$.

Phase2

$\mathcal{B}$ responds to $\mathcal{A}$'s extraction query T in the same way as in **Phase1**. $\mathcal{B}$ responds to $\mathcal{A}$'s decryption query as follows. The followings are done in a sequential way.

Step1

If $\mathsf{Verify}(vk, c_1||c_2||T||\mathsf{ID}||, \sigma) = \mathtt{reject}$, then $\mathcal{B}$ returns $\perp$ and skips **step2**$\sim$**4**.

Step2

If $vk = vk^*$, then $\mathcal{B}$ stops the simulation and outputs a random bit b'.

Step3

If $(c_1, T) = (c_1^*, T^*)$, then $\mathcal{B}$ returns $\perp$ and skips **step4**.

Step4

$\mathcal{B}$ responds in the same way as in **Phase1**.

Guess

If $\mathcal{A}$ outputs a bit, then $\mathcal{B}$ outputs a same bit as its guess.

We examine the $\mathcal{B}$'s simulation of the response to decryption queries in **Phase2**. In the case of $\mathsf{Verify} = \mathtt{reject}$ in **Step1**, $\mathcal{B}$ returns $\perp$ in the same way as in our decryption algorithm, and then it perfectly simulates the challenger in IND-hID-CCA$_{\mathsf{TS}}$ game. In the case of $vk = vk^*$ in **Step2**, the event Forge occurs. In the case of $(c_1, T) = (c_1^*, T^*)$ in **Step3**, since c_1 equals to c_1^*, the decryption of c_1 is $M_0 = (m_0 \oplus r||vk^*)$ or $M_1 = (m_1 \oplus r||vk^*)$. However, since $vk \neq vk^*$, the decryption of c is $\perp$, and then $\mathcal{B}$ simulates perfectly. In the case of $(c_1, T) \neq (c_1^*, T^*)$, $\mathcal{B}$ can issue the valid decryption query (c_1, T) to $\mathcal{C}$.

If the event Forge does not occurs, $\mathcal{B}$ perfectly simulates the challengers in the IND-hID-CCA$_{\mathsf{CR}}$ game and wins the IND-hID-CCA game with the same probability that $\mathcal{A}$ wins the IND-hID-CCA$_{\mathsf{CR}}$ game, i.e., $\Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] = Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{CR}}} + \frac{1}{2}$. □

We see that

$$\begin{aligned}
\Pr[\mathtt{Succ}] &\geq \Pr[\mathtt{Succ} \wedge \overline{\mathtt{Forge}}] \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot \Pr[\overline{\mathtt{Forge}}] \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot (1 - \Pr[\mathtt{Forge}]) \\
&= \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] - \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] \cdot \Pr[\mathtt{Forge}] \\
&\geq \Pr[\mathtt{Succ}|\overline{\mathtt{Forge}}] - \Pr[\mathtt{Forge}],
\end{aligned}$$

then, from **Claim 3**, we have that

$$\Pr[\mathtt{Succ}] \geq Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{CR}}} + \frac{1}{2} - \Pr[\mathtt{Forge}].$$

If $Adv_{\Gamma,\mathcal{A}}^{\mathsf{IND\text{-}hID\text{-}CCA_{CR}}}$ is not negligible, $Adv_{\Pi,\mathcal{B}}^{\mathsf{IND\text{-}hID\text{-}CCA}} = |\Pr[\mathtt{Succ}] - \frac{1}{2}|$ is not negligible from **Claim 4**, and it contradicts our assumption. This completes the proof of **Theorem 2**. □

VII. Conclusion

In this paper, we introduced a notion of TRHIBE and defined IND-hID-CCA$_{\mathsf{CR}}$ security and IND-hID-CCA$_{\mathsf{TS}}$ security. Moreover, we showed a generic construction of TRHIBE in which a constructed scheme achieves those security if the primitive HIBE schemes are IND-hID-CCA secure and the primitive one-time signature scheme is OT-sEUF-CMA secure.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26330160.

REFERENCES

[1] T. May, “Timed-release crypto,” Manuscript, February 1993.

[2] A. C.-F. Chan and I. F. Blake, “Scalable, server-passive, user-anonymous timed release cryptography,” in *ICDCS 2005*. IEEE Computer Society, 2005, pp. 504–513.

[3] J. H. Cheon, N. Hopper, Y. Kim, and I. Osipkov, “Timed-release and key-insulated public key encryption,” in *FC 2006*, ser. Lecture Notes in Computer Science, G. Di Crescenzo and A. Rubin, Eds., vol. 4107. Springer-Verlag, 2006, pp. 191–205.

[4] ——, “Provably secure timed-release public key encryption,” *ACM Transactions on Information and System Security (TISSEC)*, vol. 11, no. 2, p. Article 4, 2008.

[5] J. Cathalo, B. Libert, and J.-J. Quisquater, “Efficient and non-interactive timed-release encryption,” in *ICICS 2005*, ser. Lecture Notes in Computer Science, S. Qing, W. Mao, J. Lopez, and G. Wang, Eds., vol. 3783. Springer-Verlag, 2005, pp. 291–303.

[6] T. Oshikiri and T. Saito, “Timed-release identity-based encryption,” *IPSJ Journal*, vol. 55, no. 9, pp. 1964–1970, sep 2014.

[7] Y. Dodis and J. Katz, “Chosen-ciphertext security of multiple encryption,” in *TCC 2005*, ser. Lecture Notes in Computer Science, J. Kilian, Ed., vol. 3378. Springer-Verlag, 2005, pp. 188–209.

[8] D. Boneh, X. Boyen, and E.-J. Goh, “Hierarchical identity based encryption with constant size ciphertext,” in *Proceedings of the 24th Annual International Conference on Theory and Applications of Cryptographic Techniques*, ser. EUROCRYPT’05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 440–456.

[9] A. Lewko and B. Waters, “New techniques for dual system encryption and fully secure hibe with short ciphertexts,” in *Theory of Cryptography*, ser. Lecture Notes in Computer Science, D. Micciancio, Ed. Springer Berlin Heidelberg, 2010, vol. 5978, pp. 455–479.

[10] R. C. Merkle, “A digital signature based on a conventional encryption function,” in *A Conference on the Theory and Applications of Cryptographic Techniques on Advances in Cryptology*, ser. CRYPTO ’87. London, UK, UK: Springer-Verlag, 1988, pp. 369–378. [Online]. Available: http://dl.acm.org/citation.cfm?id=646752.704751

A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms

Hayden Wimmer
College of Business
Bloomsburg University
Bloomsburg, PA USA

Loreen Powell
College of Business
Bloomsburg University
Bloomsburg, PA USA

***Abstract*—While research has been conducted in machine learning algorithms and in privacy preserving in data mining (PPDM), a gap in the literature exists which combines the aforementioned areas to determine how PPDM affects common machine learning algorithms. The aim of this research is to narrow this literature gap by investigating how a common PPDM algorithm, K-Anonymity, affects common machine learning and data mining algorithms, namely neural networks, logistic regression, decision trees, and Bayesian classifiers. This applied research reveals practical implications for applying PPDM to data mining and machine learning and serves as a critical first step learning how to apply PPDM to machine learning algorithms and the effects of PPDM on machine learning. Results indicate that certain machine learning algorithms are more suited for use with PPDM techniques.**

Keywords—Privacy Preserving; Data Mining; Machine Learning; Decision Tree; Neural Network; Logistic Regression; Bayesian Classifier

I. Introduction

Knowledge discovery in databases (KDD), or Data Mining (DM), seeks to uncover patterns and relationships contained in data. Privacy of information has come under increasing scrutiny with the advent of regulations such as HIPAA [1, 2]. Simply removing fields or obscuring the records would distort the knowledge contained within the data. This necessity led to the inception of privacy preserving in data mining, or PPDM. PPDM algorithms attempt to de-identify data while maintaining the knowledge contained within. The goal of PPDM research is minimal knowledge distortion; however, some knowledge may be lost when applying PPDM. Machine learning techniques are frequently employed in KDD, or data mining. This research aims to understand the effects of PPDM on common machine learning algorithms and serves as a first step toward mapping the effects of PPDM algorithms on machine learning algorithms. Specifically, this research compares artificial neural networks (ANN), Bayesian Classifier, Decision Stump, C4.5 Decision Tree Induction, Logistic Regression, and Classification and Regression Trees (CART). This work has practical implications for data science and analytics as applied by academics and practitioners alike. The remainder of this paper is structured as follows: section 2 provides a background of machine learning algorithms and privacy preserving in data mining, section 3 presents the methodology and results, and section 4 discusses conclusions and future directions.

II. Background

A. Neural Networks

Neural networks, artificial neural networks, ANN, or NN is a computational technique which is modeled after the human brain's neural pathways [3]. ANNs are frequently applied to pattern recognition and classification and have been applied to facial recognition [4]. An ANN has an input layer and an output layer with one or more (1...n) hidden layers. The hidden layers of the ANN apply a mathematical function to the input and are said to learn by employing techniques such as adjusting weights of the input in the hidden layer. A simple, single layer, ANN is shown as *Figure 1.* ANNs have been successfully applied to many scenarios that are of interest to data science. Examples of ANN applications include recognizing financial distress patterns [5], bankruptcy prediction [6-9], and decision support systems [10]. ANNs have been applied to classic problems such as stock price forecasting [11] and medical diagnoses [12].

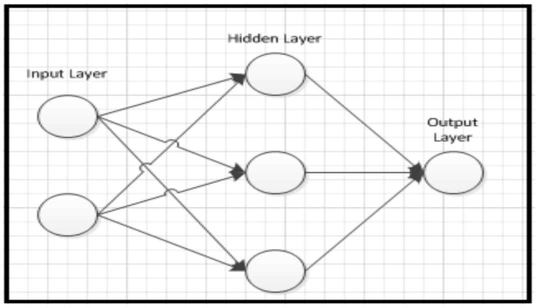


Fig. 1. A Simple Artificial Neural Network

B. Decision Trees

Decision tree algorithms are machine learning algorithms that accept data as an input and output a graph structure. Decision trees begin with a root node which branch into child nodes. A leaf node is a node with no children. Rules, as applied in expert systems, can be extracted from a decision tree. An example decision tree constructed from the classic weather data set as described by Livingston [13] is shown as *Figure 2*. The weather dataset has 14 instances and 5 features. The resulting decision tree is used to determine whether to perform a task, such as play a game, given weather conditions. One can extract rules such as *Table 1.*

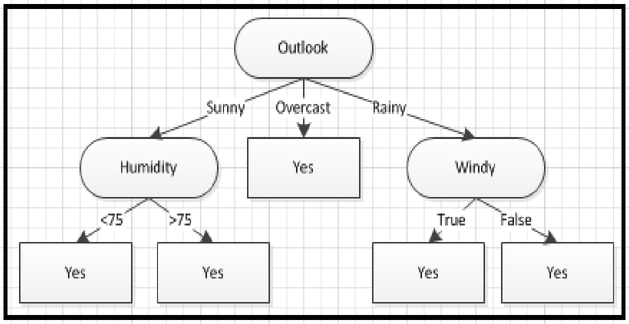
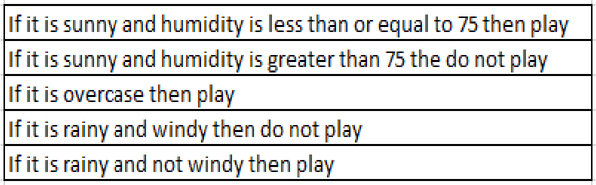


Fig. 2. A Decision Tree from Weather Data

TABLE I. RULE SET FROM DECISION TREE



If it is sunny and humidity is less than or equal to 75 then play
If it is sunny and humidity is greater than 75 the do not play
If it is overcase then play
If it is rainy and windy then do not play
If it is rainy and not windy then play

Decision tree learning algorithms include the classification and regression tree algorithm, ID3, C4.5, C5.0, CHAID, and decision stumps to name a few. Classification and regression trees, or CART, was conceived in 1984 by Breiman, et al. [14] and recursively works through data and using an index feature, the Gini index [15], and divides the data into a tree structure. ID3 [16], C4.5 [17], and C5.0 [18] are all related as C5.0 extends C4.5 and C4.5 extends ID3. ID3 and C4.5 are open source whereas C5.0 is proprietary. Like the CART algorithm, ID3 and C4.5 recursively employ a function to split the data into a tree structure; however, C4.5 and ID3 apply an entropy function which seeks to minimize the information loss occurring from each split of the data which is computed as the difference between the normalized information gains. C4.5 has been widely applied to domains such as network traffic classification[19], vehicle traffic pattern and driving behavior classification [20] patient classification [21], and organ classification [22]. CHAID is the Chi-Squared Automatic Interaction Detection algorithm and is similar to the aforementioned classification algorithms but is based in Bonferroni statistical testing [23, 24]. CHAID has a wide range of applications and has been used in financial distress classification [25].

C. *Bayesian Classifier*

Naïve Bayesian classifiers employ simple statistical assumptions to make classifications. These assumptions assist in increasing the performance of the classifier. The performance and assumptions make it an effective classifier for many applications such as junk mail filtering [26]. The Naïve Bayes classifier is considered one of the most efficient and effective classification algorithms [27]. The principle assumption made by a Naïve Bayes classifier is that all features, or independent variables, contribute equally to the target, or dependent variable. The effectiveness, regardless of the assumptions, is the optimality of classification is not necessarily related to the independence of the assumptions [28]. Despite its simplicity in its assumptions, it has been shown to outperform more powerful classifiers under many conditions which demonstrates Bayesian classifiers are a highly applicable classifier to many domains and classification problems [29]. Modern approaches include medical applications such as heart attack prediction [30], credit scoring [31], and social network analysis [32].

D. *Logistic Regression*

Logistic regression is a form of classifier that, given input independent variables, predict the target or dependent variable. Logistic regression is similar to linear regression with the exception that the target variable, or dependent variable, is categorical as opposed to continuous [33]. The dependent variable is a binary value {0, 1} frequently representing {yes, no}, {up, down}, or {good, bad}. Logistic regression has a wide range of applications such as making predictions in healthcare settings [34-36]. Logistic regression has seen modern applications in medical diagnoses [37] and in data science and analysis [38].

E. *Privacy preserving data mining and K-Anonymity*

Privacy and preserving in data mining, or PPDM, is a research stream that seeks to insert privacy into data mining while maintaining the integrity of the knowledge contained within the data [39]. The need for PPDM was emphasized in the late 1990s when the medical history of the governor of Massachusetts was uncovered by reassembling public census records with public medical data. This process, known as re-identification, detailed the need for anonymization when sharing medical data and hence the Datafly algorithm was introduced [40-42]. Some of the primary PPDM techniques include data perturbation, randomized response, condensation [43], data and rule hiding[44, 45], cryptography, noise adding, blocking, generative based, and sanitization based [46], and differential privacy [47, 48] to name a few.

Among the aforementioned PPDM techniques are algorithms that transform data to meet a standard, k-Anonymity. K-Anonymity states that each record must not be distinguishable from k respondents. In data records, there are attributes that uniquely identify individuals, such as a social security number. These attributes are considered identifier attributes. In addition to identifier attributes, there are attributes that, when combined with other attributes, uniquely identify individuals. These are referred to as quasi-identifiers. In 2000, it was found that 87% of individuals could be uniquely identified by the quasi-identifiers date of birth, zip code, and gender [49]. K-anonymity requires that, within a table, a set of quasi-identifying attributes must appear at least *k* times. For example, given the set or quasi-identifiers *S = {date of birth, zip code, gender}* and *t* is an instance in S such that *t= {2/20/1967, 98520, M}*, and *k=2,* then a minimum of 2 occurrences of tuple *t* is required for k-anonymity.

III. METHOD

A. *Framework Experiment*

The aim of this applied research is to begin mapping the effects of PPDM techniques on machine learning algorithms. First, a data file is read and classified with a machine learning algorithm. Second, the same data file is read, a privacy framework applied, and the same machine learning algorithm is applied. The general framework is detailed in *Figure 3.*

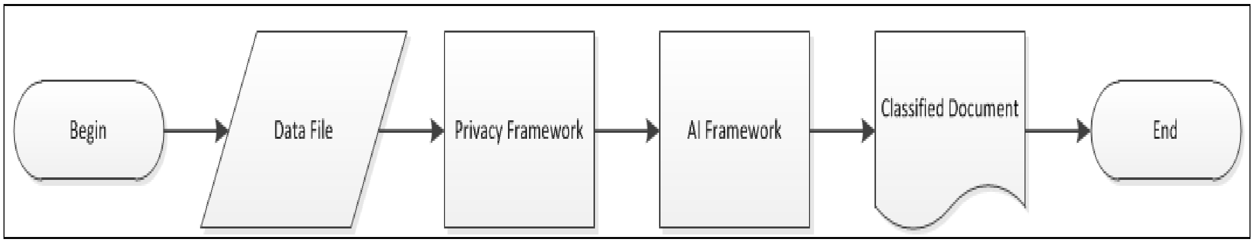


Fig. 3. General Framework for PPDM and ML

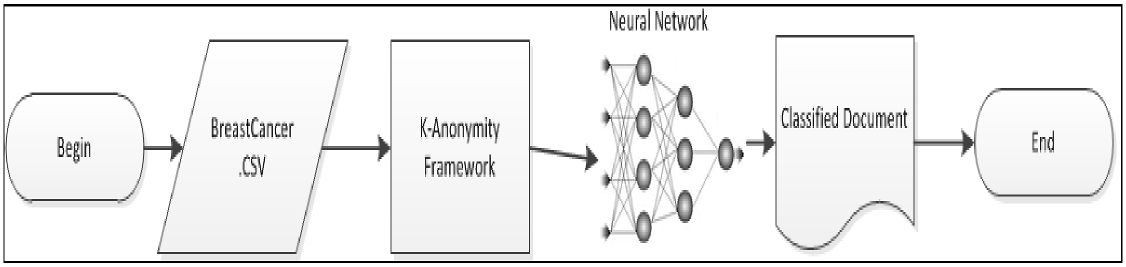


Fig. 4. Flow for ANN

Figure 4 shows the specific framework. In this work, k-anonymity is employed as the PPDM technique. Once k-anonymity with *k=2* is applied to the input dataset, the resulting anonymized dataset becomes input for 6 machine learning algorithms: 1)artificial neural network (ANN), 2) C4.5 decision tree, 3) decision stump algorithm, 4) classification and regression Tree, 5) Naïve Bayes classifier, and 6) a logistic regression. The flow for the ANN is shown as *Figure 4*.

B. Data Pre-Processing

The framework was applied to 3 separate datasets of differing size and attributes. The first dataset was extracted from [42]. The original Sweeney dataset had 5 features and 5 instances; however, an artificially generated first and last name and a target for classification were generated altering the dataset to 8 features and 5 instances. The second dataset was retrieved from the UCI machine learning repository [50] and was cited in [51] and in [52]. This dataset will be hereafter referred to as the cancer dataset. First, instances missing data were removed. Next, features for first, last, and middle name were added and randomly generated from a random name generator. The resulting dataset had 14 features and 699 instances. The final dataset was also extracted from the UCI machine learning repository and was originally extracted from census data and had 299999 instances and 6 features with 1 being an identifier and 1 being a target. Instances with missing data were removed. This dataset is named income.

C. Algorithm Parameters

All algorithms were run on a dedicated Windows 8 machine with an Intel i3 2.30GHZ processor and 8GB of physical memory. Machine learning algorithms examined include artificial neural networks, naïve Bayesian classifier, logistic regression, Decision Stump, Classification and Regression Trees (CART), and C4.5 decision tree induction. All algorithms were trained with 10 fold cross-validation. The artificial neural network was set to train through a maximum of 100 epochs and the number of hidden layers was set to a maximum of the average of the number of classes and the number of attributes. Back propagation was employed by the classifier and, if numeric, nodes are non-threshold linear units, otherwise, they are sigmoid. The naïve Bayes classifier employed is based on [53] where estimator values are chosen based on the training data. Logistic regression used multinomial regression paired with a ridge estimator as detailed in [54]. The Decision Stump algorithm is based on mean-squared error and information entropy. The minimum number of instances for a leaf was set as 1. The CART algorithm is based on [14] and implemented minimal cost-complexity pruning and 100% of the data was available to the 10 fold cross-validation process. The C4.5 decision tree algorithm [16, 17]was set to a minimum of 5 objects per leaf.

IV. Results

The results presented in Tables 2, 3, and 4 correspond to the datasets Sweeney, Cancer, and Income respectively. In reviewing the results it is necessary to consider the resulting confusion matrix from the classifier. A confusion matrix details how instances are classified. Specifically, the confusion matrix details true positives or instances correctly classified as positive, false positives or instances incorrectly classified as positive, false negatives or instances that were incorrectly classified as negative and true negatives or instances that were correctly classified as negative. In our example, the resulting confusion matrix is a 2x2 matrix and can be interpreted as *Figure 5.*

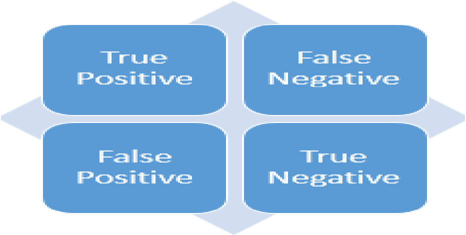


Fig. 5. Interpreting a Confusion Matrix

In each of the tables 2, 3, and 4 the resulting classification accuracy of the machine learning algorithm and confusion matrix is presented both before the PPDM technique (k-anonymity *k=2*) is applied. Any PPDM technique will make changes to the data and therefore the knowledge contained within. It is also theorized that the larger the dataset the less individual changes to individual instances a PPDM algorithm will have to introduce into the data. For example, with only 2 instances both instances will need to be changed for de-identification whereas a large dataset with a million instances will likely not have to change each individual instance to de-identify the data.

Minimizing the change in the hidden knowledge is the goal of any PPDM technique. The results show some machine learning algorithms performing better after the PPDM technique was applied which is suspect as it is natural for the knowledge to degrade after the PPDM technique.

The ANN was susceptible to this phenomenon with it improving for both the Sweeney and Income datasets and only slightly degrading for the Income dataset. In the cancer dataset prior to anonymization the resulting confusion matrix showed no true negatives or false negatives indicating that the classifier classified all data as positive with a high error rate (34.48%) which also happened in the Sweeney dataset after the PPDM technique. The confusion matrices of the income dataset reveal a decrease in true positives and false positives but an increase in false negatives and true negatives. Based on the aforementioned figures artificial neural networks did not perform well on the datasets when combined with PPDM.

The performance of the C4.5 decision tree algorithm was unchanged post PPDM for the Sweeney dataset and had a decrease in classification accuracy for the cancer and income datasets. The confusion matrix was unchanged for the Sweeney dataset but only the number of false positives and true positives changed for the cancer dataset indicating a shift in records being incorrectly classified as positive. This was 23 instances out of 699 or 3.1%. The results were different for the income dataset with a shift with an increase in false negatives and true negatives. There were 246432 correctly classified instances before and 244312 correctly classified instances after PPDM for a change in less than 1%. This indicates that C4.5 performed well with the datasets and k-anonymity.

The decision stump algorithm remained unchanged post-PPDM for the Sweeney dataset and decreased for the cancer dataset. There was a decrease in true positives and true negatives and corresponding increases in false positives and false negatives. In the case of the income dataset the decision stump simply classified everything as negative. Based on this, there are concerns with the performance of the decision stump algorithm with the datasets and PPDM technique.

The Classification and Regression Tree, or CART, algorithm was unable to make any classifications in the Sweeney dataset due to the small size. CART demonstrated a large improvement in classification accuracy in the cancer dataset indicating a potential concern. Applied to the income dataset there was a decrease in classification accuracy (82.16 to 81.43) and there was a reduction in true positives but an increase in true negatives. This indicates CART has potential on only 1 of the 3 datasets.

Naïve Bayes showed decreases in all classification accuracies for the 3 datasets which, as previously stated, is to be expected. The confusion matrices showed a decrease in true positives and true negatives for the Sweeney and cancer datasets with a decrease in true positives and increase in true negatives for the income dataset. The changes in the cancer dataset were small with a true positive reduction of less than 1% and a 14% reduction in true negatives. In the income dataset there was a 76% reduction in true positives with only a 4% increase in true negatives.

Logistic regression was unchanged for the Sweeney dataset and classification accuracy decreased for both the cancer and income datasets. Logistic regression demonstrated a decrease in true positives and true negatives in both the cancer and income datasets. True positive reduction was 2.4% and 95% and true negative reduction was 12% and 2%.

The aforementioned results are open to interpretation and can be interpreted differently based on the objectives of the PPDM and machine learning project. The results indicate that C4.5 performs best with K-anonymity, with Naïve Bayes second, and logistic regression third. The remaining 3 approaches (ANN, Decision Stump, and CART) seemed to be problematic among the 3 datasets and, while pairing any machine learning algorithm with PPDM care is critical, pairing k-anonymity with ANN, decision stump, and CART should be performed with additional cautionary measures.

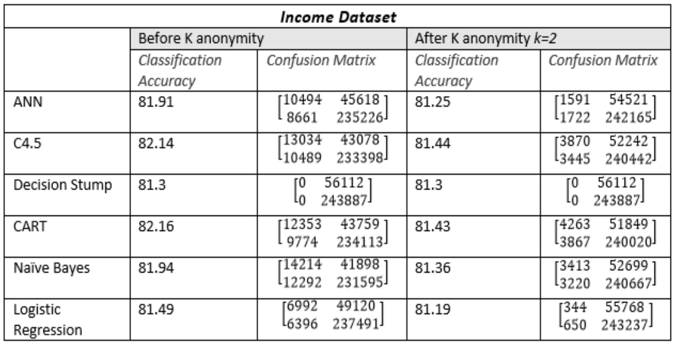
TABLE II. RESULTS FROM SWEENEY DATASET

	Sweeny Dataset			
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	60	$\begin{bmatrix}0 & 2\\0 & 3\end{bmatrix}$
C4.5	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\2 & 1\end{bmatrix}$
Decision Stump	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$
CART	NA	NA	NA	NA
Naïve Bayes	80	$\begin{bmatrix}1 & 1\\0 & 3\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$
Logistic Regression	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$

TABLE III. RESULTS FROM CANCER DATASET

	Cancer Dataset			
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	65.52	$\begin{bmatrix}458 & 0\\241 & 0\end{bmatrix}$	90.96	$\begin{bmatrix}433 & 25\\38 & 201\end{bmatrix}$
C4.5	94.42	$\begin{bmatrix}436 & 22\\17 & 224\end{bmatrix}$	91.1	$\begin{bmatrix}436 & 22\\40 & 199\end{bmatrix}$
Decision Stump	92.41	$\begin{bmatrix}417 & 41\\12 & 229\end{bmatrix}$	87.52	$\begin{bmatrix}446 & 12\\75 & 164\end{bmatrix}$
CART	64.95	$\begin{bmatrix}454 & 4\\221 & 0\end{bmatrix}$	90.96	$\begin{bmatrix}435 & 23\\40 & 199\end{bmatrix}$
Naïve Bayes	95.99	$\begin{bmatrix}436 & 22\\6 & 235\end{bmatrix}$	91.25	$\begin{bmatrix}433 & 25\\36 & 203\end{bmatrix}$
Logistic Regression	96.14	$\begin{bmatrix}444 & 16\\11 & 230\end{bmatrix}$	91.23	$\begin{bmatrix}433 & 25\\36 & 203\end{bmatrix}$

TABLE IV. RESULTS FROM INCOME DATASET



Income Dataset				
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	81.91	$\begin{bmatrix}10494 & 45618\\ 8661 & 235226\end{bmatrix}$	81.25	$\begin{bmatrix}1591 & 54521\\ 1722 & 242165\end{bmatrix}$
C4.5	82.14	$\begin{bmatrix}13034 & 43078\\ 10489 & 233398\end{bmatrix}$	81.44	$\begin{bmatrix}3870 & 52242\\ 3445 & 240442\end{bmatrix}$
Decision Stump	81.3	$\begin{bmatrix}0 & 56112\\ 0 & 243887\end{bmatrix}$	81.3	$\begin{bmatrix}0 & 56112\\ 0 & 243887\end{bmatrix}$
CART	82.16	$\begin{bmatrix}12353 & 43759\\ 9774 & 234113\end{bmatrix}$	81.43	$\begin{bmatrix}4263 & 51849\\ 3867 & 240020\end{bmatrix}$
Naïve Bayes	81.94	$\begin{bmatrix}14214 & 41898\\ 12292 & 231595\end{bmatrix}$	81.36	$\begin{bmatrix}3413 & 52699\\ 3220 & 240667\end{bmatrix}$
Logistic Regression	81.49	$\begin{bmatrix}6992 & 49120\\ 6396 & 237491\end{bmatrix}$	81.19	$\begin{bmatrix}344 & 55768\\ 650 & 243237\end{bmatrix}$

V. DISCUSSIONS AND FUTURE DIRECTIONS

The aim of the research presented in this work is to developing an understanding of the effects of PPDM techniques on machine learning algorithms. Specifically, the effects of K-Anonymity were tested against artificial neural networks (ANN), Bayesian classifiers, Decision Stump algorithm. C4.5 Decision tree induction, logistic regression, and classification and regression tree (CART) algorithm. The machine learning algorithms were tested on datasets of differing sizes and features before and after a privacy preserving data mining algorithm was applied. Results indicate that certain machine learning algorithms are more suited to use with PPDM techniques than others. This research opens the possibility for other researchers to continue and contribute by applying different PPDM techniques with machine learning algorithms. Limitations include a lack of additional datasets with a higher number of features, theoretical justification on performance, and a lack of other PPDM and machine learning algorithms. Future work will include more extensive datasets, a deeper theoretical justification, and comparing additional PPDM techniques and machine learning algorithms, specifically frequent itemset hiding and the A-priori algorithm.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of [1] G. J. Annas, "HIPAA regulations-a new era of medical-record privacy?," New England Journal of Medicine, vol. 348, pp. 1486-1490, 2003.

[2] C. F. D. Control and Prevention, "HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services," MMWR: Morbidity and Mortality Weekly Report, vol. 52, pp. 1-17, 19, 2003.

[3] M. T. Hagan, H. B. Demuth, and M. H. Beale, Neural network design: Pws Pub. Boston, 1996.

[4] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, pp. 23-38, 1998.

[5] P. K. Coats and L. F. Fant, "Recognizing financial distress patterns using a neural network tool," Financial Management, pp. 142-155, 1993.

[6] K. C. Lee, I. Han, and Y. Kwon, "Hybrid Neural Network Models for Bankruptcy Prediction," Decision Support Systems, vol. 18, pp. 63-72, 1996.

[7] K. Tam, M, "Predicting Bank Failures: A Neural Network Approach," Applied Artificial Intelligence: An International Journal, vol. 4, pp. 265-282, 1990.

[8] K. Tam and M. Kiang, "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," Management Science, vol. 38, pp. 926-948, 1992.

[9] R. Wilson and R. Sharda, "Bankruptcy Prediction Using Neural Networks," Decision Support Systems, vol. 11, pp. 545-557, 1994.

[10] N. Kumar, R. Krovi, and B. Rajagopalan, "Financial decision support with hybrid genetic and neural based modeling tools," European Journal of Operational Research, vol. 103, pp. 339-349, 1997.

[11] E. Hadavandi, H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," Knowledge-Based Systems, vol. 23, pp. 800-808, 2010.

[12] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, vol. 8, pp. 150-154, 2011.

[13] F. Livingston, "Implementation of Breiman’s random forest machine learning algorithm," ECE591Q Machine Learning Journal Paper, 2005.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and regression trees: CRC press, 1984.

[15] R. I. Lerman and S. Yitzhaki, "A Note on the Calculation and Interpretation of the Gini Index," Economics Letters, vol. 15, pp. 363-368, 1984.

[16] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, pp. 81-106, 1986.

[17] J. R. Quinlan, C4. 5: programs for machine learning vol. 1: Morgan kaufmann, 1993.

[18] J. R. Quinlan. (2012). C5.0: An Informal Tutorial.

[19] Y. Zhang, H. Wang, and S. Cheng, "A method for real-time peer-to-peer traffic classification based on C4. 5," in Communication Technology (ICCT), 2010 12th IEEE International Conference on, 2010, pp. 1192-1195.

[20] Z.-W. Yuan and Y. Dong, "Research the association of dangerous driving behavior and traffic congestion based on C4. 5 algorithm," Computer, Intelligent Computing and Education Technology, p. 403, 2014.

[21] A. G. Karegowda, V. Punya, M. Jayaram, and A. Manjunath, "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4. 5," International Journal of Computer Applications, vol. 45, 2012.

[22] M. K. Ross, K.-W. Lin, K. Truong, A. Kumar, and M. Conway, "Text categorization of Heart, Lung, and Blood studies in the Database of Genotypes and phenotypes (dbGap) Utilizing n-grams and Metadata Features," Biomedical informatics insights, vol. 6, p. 35, 2013.

[23] E. Antipov and E. Pokryshevskaya, "Applying CHAID for logistic regression diagnostics and classification accuracy improvement," Journal of Targeting, Measurement and Analysis for Marketing, vol. 18, pp. 109-117, 2010.

[24] J. Magidson, "The chaid approach to segmentation modeling: Chi-squared automatic interaction detection," Advanced methods of marketing research, pp. 118-159, 1994.

[25] N. Ozgulbas and A. S. Koyuncugil, "Developing Road Maps for Financial Decision Making by CHAID Decision Tree: CHAID Decision Tree Application," in Information Management and Engineering, 2009. ICIME '09. International Conference on, 2009, pp. 723-727.

[26] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in Learning for Text Categorization: Papers from the 1998 workshop, 1998, pp. 98-105.

[27] H. Zhang, "The optimality of naive Bayes," A A, vol. 1, p. 3, 2004.

[28] I. Rish, "An empirical study of the naive Bayes classifier," presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001.

[29] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine learning, vol. 29, pp. 103-130, 1997.

[30] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, pp. 250-255, 2010.

[31] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," Expert Systems with Applications, vol. 37, pp. 534-545, 2010.

[32] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in Social network data analytics, ed: Springer, 2011, pp. 243-275.

[33] D. W. Hosmer Jr and S. Lemeshow, Applied logistic regression: John Wiley & Sons, 2004.

[34] J. J. Tomaszewski, R. G. Uzzo, N. Kocher, T. Li, B. Manley, R. Mehrazin, et al., "Patients with anatomically “simple” renal masses are more likely to be placed on active surveillance than those with anatomically “complex” lesions," in Urologic Oncology: Seminars and Original Investigations, 2014.

[35] A. C. Davis, G. Watson, N. Pourat, G. F. Kominski, and D. H. Roby, "Disparities in CD4 Monitoring among HIV-Positive Medicaid Beneficiaries: Evidence of Differential Treatment at the Point of Care," in Open Forum Infectious Diseases, 2014, p. ofu042.

[36] E. Dahlén, C. Almqvist, A. Bergström, B. Wettermark, and I. Kull, "Factors associated with concordance between parental‐reported use and dispensed asthma drugs in adolescents: findings from the BAMSE birth cohort," Pharmacoepidemiology and Drug Safety, 2014.

[37] D. Timmerman, B. Van Calster, A. C. Testa, S. Guerriero, D. Fischerova, A. Lissoni, et al., "Ovarian cancer prediction in adnexal masses using ultrasound‐based logistic regression models: a temporal and external validation study by the IOTA group," Ultrasound in obstetrics & gynecology, vol. 36, pp. 226-234, 2010.

[38] J. Sall, A. Lehman, M. L. Stephens, and L. Creighton, JMP start statistics: a guide to statistics and data analysis using JMP: SAS Institute, 2012.

[39] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Record, vol. 29, pp. 439-450, 2000.

[40] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the Datafly System," in Proceedings of the AMIA Annual Fall Symposium, 1997, p. 51.

[41] L. Sweeney, "Datafly: a system for providing anonymity in medical data," Database Security, XI: Status and Prospects, 1998.

[42] L. Sweeney, "Computational disclosure control for medical microdata: The Datafly system," in Record Linkage Techniques 1997: Proceedings of an International Workshop and Exposition, 1997, pp. 442-453.

[43] G. Nayak and S. Devi, "a survey on privacy preserving data mining: approaches and Techniques," International Journal of Engineering Science and Technology (IJEST), vol. 3, pp. 2117-2133, 2011.

[44] V. S. Verykios, "Association rule hiding methods," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, pp. 28-36, 2013.

[45] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM Sigmod Record, vol. 33, pp. 50-57, 2004.

[46] R. Natarajan, R. Sugumar, M. Mahendran, and K. Anbazhagan, "A survey on Privacy Preserving Data Mining," International Journal on Advanced Research in Computer and Communications Engineering, vol. 1, 2012.

[47] C. Dwork, "Differential privacy: A survey of results," in Theory and Applications of Models of Computation, ed: Springer, 2008, pp. 1-19.

[48] C. Dwork, "Differential privacy," in Automata, languages and programming, ed: Springer, 2006, pp. 1-12.

[49] L. Sweeney, "Simple demographics often identify people uniquely," Health (San Francisco), pp. 1-34, 2000.

[50] (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/

[51] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," Proceedings of the national academy of sciences, vol. 87, pp. 9193-9196, 1990.

[52] J. Zhang, "Selecting typical instances in instance-based learning," 1992.

[53] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp. 338-345.

[54] S. Le Cessie and J. Van Houwelingen, "Ridge estimators in logistic regression," Applied statistics, pp. 191-201, 1992.