

ISSN : 2394 - 2231

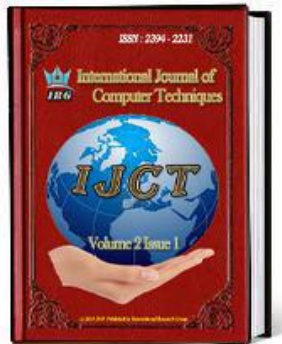


# International Journal of Computer Techniques



Volume 2 Issue 5

## **IJCT Journal**



The "**International Journal of Computer Techniques**" is a leading international journals for publication of new ideas, the state of the art research results and fundamental advances in all aspects of computer science and engineering. IJCT is a scholarly open access, peer reviewed international journal with a primary objective to provide the academic community and industry for the submission of half of original research and applications related to Computer Science and Engineering

Submitted papers will be reviewed by Technical Committees of the Journal. All submitted articles should report original, experimental or theoretical, and will be peer-reviewed. Articles submitted to the journal should meet these criteria and must not be under consideration for publication elsewhere. Manuscripts should follow the style of the journal and are subject to both review and editing.

### **Why IJCT:**

- IJCT is indexed in Google Scholar, OAJI Index, Slideshare, Scribd, CiteFactor, Academia, Issuu, Dostosc and many more.
- IJCT is an International, Peer reviewed, Open Access Journal.
- Paper publishing process is relatively fast and easy.
- Open access journal database for high visibility and promotion of your articles.
- Authors can download their full length paper at any time.
- Author's queries shall be resolved within 24 hours of time.

All manuscripts are subject to rapid peer review. Those of high quality (not previously published and not under consideration for publication in another journal) will be published without delay.

**Submit Your article:** [editorijctjournal@gmail.com](mailto:editorijctjournal@gmail.com)

**Frequency:** 6 issues per year

**ISSN :** 2394-2231

**Subject :** Computer Science Techniques

**For Published By:** International Research Group™

**For Board Members:** [editor@ijctjournal.org](mailto:editor@ijctjournal.org)



# I J C T

International Journal  
of Computer TechniQues!..

A Peer-reviewed Scholarly Research Publishing Journal

ISSN : 2394-2231



## Board Members:



- **Prof.Mr.Galal Ali Hassaan**  
- Editor - in- Chief of IJCT  
- (Cairo University), MASME,MAMSE. Egypt.  
- Department of Mechanical Design and Production,  
Faculty of Engineering, Cairo University, Giza, Egypt.
- **Dr. Saurabh Patel**  
VBS Purvanchal University, Jaunpur.
- **Prof. Jignesh G. Bhatt**  
Dharmsinh Desai University, Nadiad,Gujarat, India.
- **Dr.N.satyan**  
Associate Professor of CSE in KL University. A.P
- **Dr. Pallamreddy**  
QIS College of Engineering and Technology, AP., University of Mysore
- **Dr. Sanjay**  
L.N.C.T Bhopal.
- **Dr. K Rajesh Kumar**  
Vijayawada, Andhra Pradesh, India.
- **Dr. P. Rameshkrishna**  
Dept of Information Technology, Hindustan University, Chennai
- **Dr. Vijay Mankar**  
Dept of Electronics Telecommunication Eng, Maharashtra.
- **Dr.M.Hanumanthappa**  
Dept of Studies in Computer Science, University of Mysore
- **Dr.Venkat Rai**  
Hyderabad Institute of Technology Management., University of Mysore
- **Dr. V.Balamurugan**  
,Einstein College of Engg,Tirunelveli,India.
- **Dr.J.Hanumanthappa**  
Dept of Studies in Computer Science, University of Mysore
- **Dr.G. Meyyappan**  
Alagappa University, Karaikudi , India.
- **Dr. DonFord**  
University of North Carolina at Charlotte, USA.
- **Dr. Saurabh Pal**  
Department of MCA,VBS Purvanchal University, Jaunpur.
- **Prof. Jignesh G. Bhatt**  
Dharmsinh Desai University, Nadiad,Gujarat, India.
- **Dr.T.V.Sai Krishna**  
QIS College of Engineering Technology, Ongole, AP.
- **Dr.s.satyanarayana**  
Associate Professor of CSE in KL University. A.P
- **Dr. Sudarson Jena**  
GITAM University ,AP.



- **Dr. S.N.Singh**  
National institute of Technology, Jamshedpur.
- **Dr P.Umamaheswari**  
Anna University, Chennai.



## IJCT Indexing:

The articles accepted and published with International Journal of Computer Techniques (IJCT) are indexed with the following academic databases.





# I J C T

International Journal  
of Computer TechniQues!..

A Peer-reviewed Scholarly Research Publishing Journal

ISSN : 2394-2231



YAHOO!



CiteSeer<sup>x</sup> beta

New Jour  
• Electronic Journals & Newsletters •



Electronic  
Journals Library

[Computer Science Directory](#). We are listed under [Publication Journals](#) category

### University Affiliates






International Journal of Computer Techniques(IJCT) is endorsed and supported by the following universities. The bibliographies of the published articles of International Journal of Computer Techniques (IJCT) are available with the universities' libraries.



## Table of Contents

### International Journal of Computer Techniques

**IJCT -Volume 2, Issue 5 / Sep - Oct - 2015**

S.No	Title/Author Name	Page No
<b>1</b>	<b>Learning Media Introduction of Yogyakarta Culture For Early Childhood 2-3 Years</b> -  Lisna Zahrotun, Yana Hendriana, Damba Saputra	<b>1-5</b>
<b>2</b>	<b>Image Compression Using Hybrid Combinations of DCT SVD and RLE</b> -  Raghavendra.M.J, Dr.Prasantha .H.S , Dr.S.Sandya	<b>6-19</b>
<b>3</b>	<b>A Survey of ETL Tools</b> -  Mr. Nilesh Mali, Mr.SachinBojewar	<b>20-26</b>
<b>4</b>	<b>Towards measuring learner?s concentration in E-learning systems</b> -  Saoutarrih Marouane, Sedrat Najlaa, Tahiri Abderrahim , Elkadiri Kamal Eddine	<b>27-29</b>
<b>5</b>	<b>A Study on Behavioral Malware Detection by Using Delay Tolerant Networks</b> -  K.Ravikumar,V. Vinothkumar	<b>30-33</b>
<b>6</b>	<b>Sybil Belief: A Semi- Creation New Approach for Structure ?Based Sybil Detection</b> -  K.Ravikumar, B.Selvam	<b>34-36</b>
<b>7</b>	<b>Multi Agent System in Distributed Agent Network</b> -  K.Ravikumar, A. Surendar	<b>37-39</b>
<b>8</b>	<b>Analysis to Online Stock for Decision Approach of Investor</b> -  G.Magesh , S. Saradha	<b>40-42</b>





<b>9</b>	<b>Cluster Ensemble Approach for Clustering Mixed Data</b> - 🧑 Honarine Mutazinda A, Mary Sowjanya , O.Mrudula	<b>43-51</b>
<b>10</b>	<b>An Improved Collaborative Filtering Algorithm Based on Tags and User Ratings</b> - 🧑 CaiyunGuo, HuijinWang	<b>52-60</b>
<b>11</b>	<b>Performance of data mining algorithms in unauthorized intrusion detection systems in computer networks</b> - 🧑 Hadi Ghadimkhani, Ali Habiboghli, Rouhollah Mostafaei	<b>61-66</b>
<b>12</b>	<b>Vibration Analysis of a Horizontal Washing Machine, Part IV: Optimal Damped Vibration Absorber</b> - 🧑 Galal Ali Hassaan	<b>67-71</b>
<b>13</b>	<b>Prediction of Heart Disease Using Enhanced Association Rule Based Algorithm</b> - 🧑 Karandeep Kaur, Ms. Poonamdeep Kaur, Ms. Lovepreet Kaur	<b>72-76</b>
<b>14</b>	<b>Enhancing Security in Wireless Sensor Network Using Load Balanced Data Aggregation Tree Approach</b> - 🧑 A.Senthilkumar, K. Madhurabhasini	<b>77-80</b>
<b>15</b>	<b>Online Shopping Product Aspect and Ranking Using Support Vector Machine Algorithm</b> - 🧑 R. Bharathi	<b>81-83</b>
<b>15</b>	<b>Online Shopping Product Aspect and Ranking Using Support Vector Machine Algorithm</b> - 🧑 R. Bharathi	<b>81-83</b>
<b>16</b>	<b>Resilient Key Sharing Approach Based on Multimode Authentication Scheme</b> - 🧑 A.Senthilkumar , R.Divya	<b>84-88</b>
<b>17</b>	<b>Secure Multiparty Computation on Multiple Clouds</b> - 🧑 K.Ravikumar, S. Thamizharasi	<b>89-92</b>






# I J C T

International Journal  
of Computer TechniQues!..

*A Peer-reviewed Scholarly Research Publishing Journal*

ISSN : 2394-2231



<b>18</b>	<b>A Survey on Log Mining: A Data Mining Approach for Intrusion Detection</b> -  Smita P. Bhapkar, Shubhangi S. Dhamane, Yogita S. Kandekar, Khushbu S. Lodha	<b>93-97</b>
<b>19</b>	<b>Bluetooth Message Hopping Chat Application</b> -  Kirti Karande, Ibrahim Shaikh, Tanzeel Shaikh, Hardik Vaghela	<b>98-102</b>
<b>20</b>	<b>Urbain Traffic Congestion Estimating Using Simplified CRONOS Model: Algorithm and Implementation</b> -  Abdallah Lakhouili , Hicham Medromi , El Hassan Essoufi	<b>103-110</b>

# Learning Media Introduction of Yogyakarta Culture For Early Childhood 2-3 Years

Lisna Zahrotun\*, Yana Hendriana\*\*, Damba Saputra\*\*\*

\*(Informatics Department, Ahmad Dahlan University, Yogyakarta, Indonesia)

\*\* (Informatics Department, Ahmad Dahlan University, Yogyakarta, Indonesia)

\*\*\* (Informatics Department, Ahmad Dahlan University, Yogyakarta, Indonesia)

\*\*\*\*\*

## Abstract:

Early Childhood Education (ECE) is very important to do well in a home environment as well as in the educational environment of pre-school. Education pre-school children at age 2 to 3 years is the formation of character, where education is stressed here in the picture, sound and movement combined with an attractive shape and color. At this age children begin to recognize the objects around it. The introduction of culture is also one of the curriculum in pre-school education.

To help children recognize the surrounding culture then designed introduction to the culture medium of learning for children aged 2 to 3 years. Where learning media aims to help children in recognizing culture.

In this study generated media that can provide an alternative to teachers in pre-school learning environment. In addition the children will be easier to recognize the culture while also able to train the child's motoricskill in moving the mouse and clicking on the image. Children's learning also feel happy and not bored.

Keywords —**Learning Media, Motoric Skills, Culture.**

\*\*\*\*\*

## I. INTRODUCTION

UNESCO argued that education should be put on the four pillars, namely learning to know, learning to do, learning to live together, and learning to be [1][2]. In the learning process, especially for children is learning that spawned a pleasant atmosphere. The pictures and sounds that emerge will make the children do not get bored quickly, so as to stimulate learning in children [3].

While the child's development with regard to the overall personality of the child, because of personality forming an integrated whole. In general it can be distinguished some aspects of child development. Aspects of early childhood development of children of 2-3 years age group consisting of [4]:

- 1) Aspect of moral and religious values

- 2) Aspect of social, emotional and self-reliance
- 3) Aspect of language
- 4) Aspect of Cognitive
- 5) Aspect of physical / motoric
- 6) Aspect of art

The world's children are playing. children will learn something new and doing various activities that are useful for the development itself, both from the aspect of cognitive, psychomotor, affective and spiritual [5]. Thus, in early childhood learning must also be within the context of learning and play.

Multimedia is one of the applications that are used as a medium of learning for childhood. Learning media based on multimedia in the classroom is developed on the basis of the assumption that the communication process in an active learning approach (active learning) can

strengthen and expedite the stimulus and the response of the students in the learning [1].

Interactive learning is the media that is designed for students to learn independently, actively and controlled. Multimedia is a very complex medium that combines some elements of the media that involve text, graphics, images, photographs, audio, video, and animation are integrated. Excess multimedia learning, among other [6]:

- 1) Enlarge objects very small and not visible to the eye.
- 2) Minimize the very large objects, which is not possible in the school.
- 3) Presenting objects or events are complex, complicated, and takes place sooner or later.
- 4) Presenting a distant object or event.
- 5) Presenting dangerous objects or events.
- 6) Provide convenience and practical learning.
- 7) Make the child active and independent learning.
- 8) Potential passion.
- 9) Allow more direct interaction between students with the environment and reality.
- 10) Allow the students to learn on their own according to their ability and interest respectively.

One of the curriculum in early childhood education, 2-3 years is the introduction of the culture. To introduce the local culture may include aspects of child development would require a tool. Where these tools can enhance students' creativity in learning, especially in the introduction of the culture. The tools used in the introduction to the culture of learning is learning introduction to the culture media. Where in this learning media in addition students can learn also can play, because in this learning media are in the form of a puzzle game.

Computer gaming is a form of real-time interactive software wrapped in creatively crafted media that offers game-players engaging, goal-directed play [7]. Thus, this concept combines games and technology to provide game players with more interesting playing experience[8]. Games play a substantial role in shaping learning techniques for kids [9].

## **II. LITERATURE REVIEW**

Children today live in different cultural settings. The pre-school culture is one of them and the media culture outside the pre-school another. These cultures are in different ways characterised by opposite and often even conflicting traditions. This article shows how educators and children handle this dilemma by using interaction as a tool to bring changes into the discourse in an educational setting while making stories in the pre-school by means of the multimedia functions of the computer. The interactional processes from three observations are described. In the discussion a comparison with another study with a constructivist point of departure is made. The comparison between the two studies showed contrasting results. The use of a socio-cultural perspective in the presented project make the context and the community visible, while the other study with its underlying assumptions of individually constructed knowledge make context and community invisible [10].

Multimedia learning occurs when students build mental representations from words and pictures that are presented to them (e.g., printed text and illustrations or narration and animation). The promise of multimedia learning is that students can learn more deeply from well-designed multimedia messages consisting of words and pictures than from more traditional modes of communication involving words alone [11].

The most basic effect of presentation method concerns whether multimedia presentations are more effective than single medium presentations, and in particular, whether adding pictures to words helps students understand an explanation. The *multimedia effect* refers to the finding that students learn more deeply from a multimedia explanation presented in words and pictures than in words alone. Let's consider how this effect fares under two learning environments with printed text and illustrations on a page and with spoken text and animation on a screen [11].

The influence of aspects of home and preschool environments upon literacy and numeracy achievement at school entry and at the end of the 3rd year of school. Individuals with unexpected performance pathways (by forming demographically adjusted groups: overachieving, average, and underachieving) were identified in

order to explore the effects of the home learning environment and preschool variables on child development. Multilevel models applied to hierarchical data allow the groups that differ with regard to expected performance to be created at the child and preschool center levels. These multilevel analyses indicate powerful effects for the home learning environment and important effects of specific preschool centers at school entry. Although reduced, such effects remain several years later [12].

Most of the research and teaching has been about media and particularly about moving image media such as television and film. It have explored how these media are produced, the characteristics of media 'text', and how children and young people use and interpret them and also considered how teachers in school might teach about these media, and what happens when they do so. Inevitably, in recent years, this focus has expanded to encompass new media such as computer games and the internet. However, It continue to regard these things as media rather than as technologies. Them as ways of representing the world, and of communicating and to understand the phenomena as social and cultural processes, rather than primarily as technical ones. Technologies or machines are obviously part of the story[13].

### III. METHODOLOGY

Learning media constructed through three stages, among others:

- Planning stage;
- Design stage
- Implementation stage.

The design phase includes the design of scenarios, content design, and user interface design.

#### A. Planning Stage / Design scenario

In planning the development of instructional media introduction of this culture, it must first be made flowchart as shown in Fig. 1.

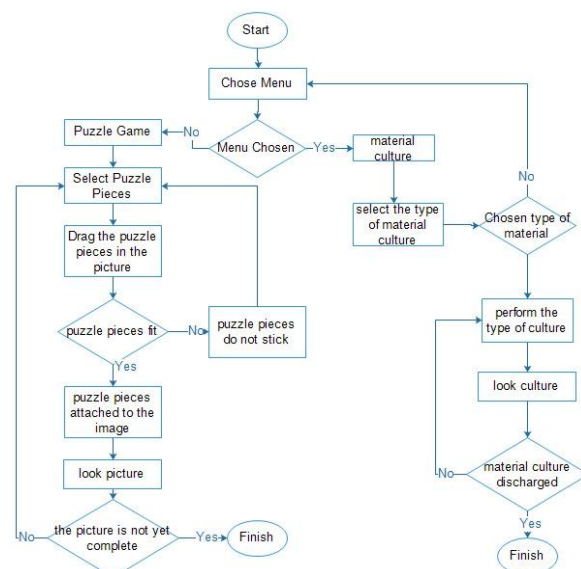


Fig. 1 Flowchart of learning media

Fig. 1 shows that in the learning media begins by selecting a menu, the menu consists of material culture and a puzzle game. If you choose the culture topic then it will go to sub-menu of the culture, in which there are four types of culture that is gamelan, wayang, clothing, traditional dance and cultural heritage. Whereas if you choose a puzzle game that will provide the puzzle pieces. Dragged puzzle pieces to be attached to the picture provided, if it does not fit the puzzle pieces will not stick if it fits then be taped and then choose the puzzle pieces to complete the picture plastered all puzzle pieces.

#### B. Design Stage

Developing of this application designed in multiple pages and layers that allow users to learn and understand the culture in an interesting way. Selection of different elements in the culture and game design materials such as text, sound, and images will determine the final outcome of this work and the extent to which applications will be most appealing to children. The elements in the design of the game can be described as follows:

##### 1) Text

Texts are used to provide further explanation about the objects displayed. The font types in this application are selected to suit the needs and are compatible with the objects described so as to make them easy to read and to provide comfortable view.

##### 2) Sound Effect



Sound effect is a supporting element which functions as a back sound so that the application looks more alive. The sound format used in this application is \*.WAV extension.

### 3) Image

The images used in this application was chosen to attract children. Image is an image gamelan, wayang image, images dance and traditional dresses and images of cultural heritage.

### C. Implementation Stage

User interface design that is used as the basis for interaction between players and gaming applications. Interface design is made by drawing a layout for each page of this instructional media. In this learning media application layout consists of a main menu, sub menus and sub-menus materials puzzle game.

## IV. RESULTS AND DISCUSSION

Opening display that will first appear when the user opens this learning media as shown in Fig. 2.



Fig. 2 Opening Display

Fig.2 Shows the opening page there is an action that serves to give effect to the full view at the time the application is opened.

### A. Sub Menu Material

Sub menumaterial contains material on the introduction of Yogyakarta culture. In this menu material provides an introduction to the types of gamelan, types of puppets, the types of batik and some cultural heritage. To get in on the types of gamelan then click on the gamelan image, as well as puppets, batik and cultural heritage as shown in fig. 3.



Fig. 3 Sub Menu Material

Detail sub menu describes Yogyakarta culture in the form of gamelan. In this instructional media students are introduced nine types of gamelan, introduced for culture of four puppets name, the custom equipment introduced custom clothing and a dagger, the cultural heritage introduced historic buildings, Detail sub menu as shown in Fig. 4.



Fig. 4 Detail Sub Menu, point a shows the different types of gamelan , point b shows the various types of puppets, point c shows custom clothing and point d shows the cultural heritage

### B. Implementation of Game Menu

The game menu contains a puzzle game that has been provided. Puzzle contains only four pieces of the image from the original image. This is because the introduction of this culture is still to early

childhood children. The big picture is the actual image. To be able to arrange the puzzle then the small image below to be dragged placed on the big picture. This is done until all installed following a complete image display game menu as shown in fig. 5.

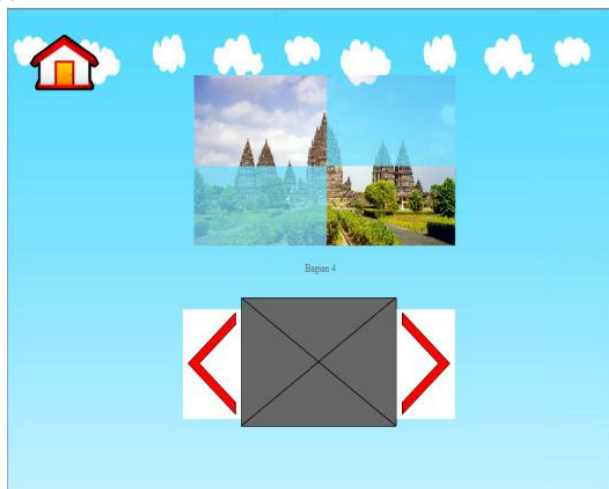


Fig. 5 Display Menu Puzzle Game

## V. CONCLUSIONS

Based on development of this learning media application, it can be concluded that:

- 1) Introduction to the material culture of the displayed image is accompanied with text and sound.
- 2) The puzzle game is produced from four pieces at random to be able to train the child's memory.
- 3) The application can be run on Windows operating systems.

## REFERENCES

- [1] Halidah, "Perancangan Aplikasi Pembelajaran Berbasis Multimedia untuk Anak Usia Dini," *J. Sist. dan Teknol. Inf.*, vol. 3, no. 1, 2014.
- [2] Mulyasa, E. Implementasi Kurikulum Tingkat Satuan Pendidikan, Kemandirian Guru dan Kepala Sekolah. Jakarta: Bumi Aksara, 2008.
- [3] S. S. Dewantik H., A. Mukminin, and E. Waluyo, "Penerapan Pembelajaran Berbasis Komputer sebagai Dasar Pengenalan Teknologi Informasi pada Guru Taman kanak-kanak di kota Semarang," *journal.unnes.ac.id*, vol. 14, no. 2, 2010.
- [4] S. Susilo, "Indikator PAUD Kelompok Usia 2-3 Tahun," 2013.
- [5] R. Apriliana, Strategi Guru dalam Mengembangkan Kreativitas pada Anak Usia Prasekolah Kelas B2 di TK 'aisyiyah Bustanul Athfal Darussalam Banyudono Dukun Magelang (semester Gasal Tahun Ajaran 2013/2014). 2014.
- [6] P. Rangsang and K. P. Adhithia, "Pemanfaatan Teknologi SMS Gateway dan Metode Forward Chaining Pada Sistem Informasi Bimbingan dan Konseling (Studi Kasus SMAK ST Thomas Aquino Mojokerto)," in *SNASTI*, 2010, pp. 16–21.

- [7] Tang, S., Hanneghan, M., *Game Content Model: An Ontology for Documenting Serious Game Design*. 2011 Developments in E-systems Engineering, pp 432–436, 2011
- [8] Annetta, L.A., Minogue, J., Holmes, S.Y., Cheng, M.T., *Investigating the impact of video games on high school students' engagement and learning about genetics*, Computers & Education 53, 74–85, 2009
- [9] Janarthanan, V. *Serious Video Games: Games for Education and Health*. Ninth International Conference on Information Technology: New Generations (ITNG), pp 875- 878, 2012
- [10] Klerfelt, Anna. Ban the computer, or make it a storytelling machine bridging the gap between the children's media culture and preschool.Scandinavian Journal of Educational Research, 2004, 48.1: 73-93.
- [11] Mayer, Richard E. The promise of multimedia learning: using the same instructional design methods across different media. *Learning and instruction*, 2003, 13.2: 125-139
- [12] Melhuish, Edward C., et al. Effects of the home learning environment and preschool center experience upon literacy and numeracy development in early primary school. *Journal of Social Issues*, 2008, 64.1: 95-114.
- [13] Buckingham, David. *Beyond technology: Children's learning in the age of digital culture*. John Wiley & Sons, 2013.

# Image Compression Using Hybrid Combinations of DCT SVD and RLE

Raghavendra.M.J<sup>1</sup>,

<sup>1</sup>Assistant Professor, Department of Telecommunication,  
P.E.S. Institute of Technology,  
Bangalore, India

Dr.Prasantha .H.S<sup>2</sup> , Dr.S.Sandya<sup>3</sup>

<sup>2,3</sup> Professor, Department of Electronics and Communication,  
Nitte Meenakshi Institute of Technology,  
Bangalore, India

\*\*\*\*\*

## Abstract:

Image Compression finds a significant place in the field of research. In this paper we are proposing a scheme for hybrid image compression which uses Discrete Cosine Transform, Singular Value Decomposition and Run length Encoding. Discrete Cosine Transform is applied to the image. Then DC-Coefficient is taken out from Discrete Cosine Transformed Matrix and stored or transmitted separately. The Discrete Cosine Transformed matrix without DC-coefficient is truncated with a threshold value. To this truncated matrix Singular Value Decomposition is applied. The matrices obtained from the Singular Value Decomposition are again truncated with suitable threshold value. Then these matrices are multiplied back. The resultant matrix is again truncated with threshold value. Then this matrix is quantized. The quantized matrix is converted into sparse matrix form. Then sparse matrix elements under goes data type conversion. The column elements of the sparse matrix are run length encoded and then compressed form of the image can be obtained. This compressed form can be stored or transmitted. An effort is also made to compare the number of memory bytes obtained in this method with the three other methods which are discussed.

**Keywords-- DCT-Discrete Cosine Transform, SVD-Singular Value Decomposition, MSE-Mean Squared Error, PSNR-Peak Signal to Noise Ratio, CR-Compression Ratio ,RLE-Run Length Encoding**

\*\*\*\*\*

## I. INTRODUCTION

There exists always demand for Image compression in the field of Multimedia. Image Compression is broadly classified into two types. They are lossless image compression techniques and lossy image compression techniques. It can be learnt that in the lossless image compression techniques the reconstructed image quality is better than the lossy image compression techniques. But when we compare with the compression ratio, lossy compression technique is better than the loss less compression technique. In this paper we are proposing hybrid image compression technique using DCT, SVD and RLE. This is a lossy compression technique.

This paper consists of seven sections. The first section deals with the introduction, the second

section deals with literature survey, the third section deals with the methodology, the fourth section deals with implementation, the fifth section deals with the results and discussions , the sixth section deals with the scope for further enhancement and the seventh section deals with the references.

## II. LITERATURE SURVEY

There are different contributions to the above discussed problem. Few papers are discussed in this section.

Raghavendra.M.J [1] and others have worked on Image Compression using DCT and SVD to achieve image compression. Prasanth.H.S and others [2] have worked on image compression using SVD. S.Sridhar and others [3] have worked on image compression using different types of

wavelets. T.D.Khadatre and others [4] have worked on compression of image using vector quantization and wavelet transform. Athira.M.S and others [5] have worked on image compression using artificial neural networks. Pallavi and others [6] have worked on image compression using Wavelets and Huffman Coding. E.Praveen Kumar and others [7] have worked on image compression using multiwavelet transforms. D.Vishnuvardhan and others [8] have worked on image compression using curvelets.

Birendrakumar Patel and others [9] have worked on image compression using Artificial Neural Networks. Sumegha.Y and others [10] have worked on fractal image compression using Discrete Cosine Transform and Discrete Wavelet Transform. Rowayda A.S [11] worked on SVD for image processing applications. K.R.Rao [12] and others have worked on DCT.

### III. METHODOLOGY

In the proposed scheme, discrete cosine transform and singular value decomposition and run length encoding are used to compress the image data.

#### Discrete Cosine Transform

Discrete cosine transform very useful in image compression. In this it will transform the energy of the signal into lower order frequency coefficients. The formula of 2-dimensional DCT for the input function  $f(x,y)$  is as follows.

$$A(u,v) = B(u)C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cos\left\{\frac{(2x+1)u\pi}{2N}\right\} \cos\left\{\frac{(2y+1)v\pi}{2N}\right\} \quad (1)$$

Where  $u = 0, 1, 2, \dots, N-1$ ,  $v = 0, 1, 2, \dots, N-1$ ,  $f(x,y)$  = input function

The inverse 2-dimensional DCT formula is as follows

$$f(x,y) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} B(u)C(v)A(u,v) \cos\left\{\frac{(2x+1)u\pi}{2N}\right\} \cos\left\{\frac{(2y+1)v\pi}{2N}\right\} \quad (2)$$

Where  $B(u) = \sqrt{1/N}$  for  $u=0$ ,  $B(u) = \sqrt{2/N}$  for  $u=1, 2, \dots, N-1$

Similarly  $C(v) = \sqrt{1/N}$  for  $v=0$ ,  $C(v) = \sqrt{2/N}$  for  $v=1, 2, \dots, N-1$

#### Singular Value Decomposition

Singular value decomposition takes rectangular matrix as input and transforms it into three matrices "U", "S" and "V". If the input rectangular matrix is "X", then the relationship between "X" and "U", "S" and "V" are  $X=U*S*V^T$ , where  $V^T$  is the transpose of the "V" matrix. If "X" matrix is of the order  $m \times n$ , then order of the "U" matrix is of  $m \times m$ , order of the "S" matrix is  $m \times n$  and the order of "V" is  $n \times n$ . The

"S" matrix is the important matrix because it has the singular values of the input matrix. The "S" matrix has only principal diagonal elements. The magnitudes of the diagonal elements are placed in decreasing order.

#### Run Length Encoding

It is a lossless compression technique. In this method number of frequently occurring symbols are counted and it encoded before the symbol. In this way it reduces the transmission bandwidth.

#### Sparse Matrix

Sparse matrix is one in which majority of the elements are zero. Since majority of the elements are zero, the sparse notation is applied to reduce the transmission bandwidth. In the sparse notation only non-zero element's row, column and value are stored.

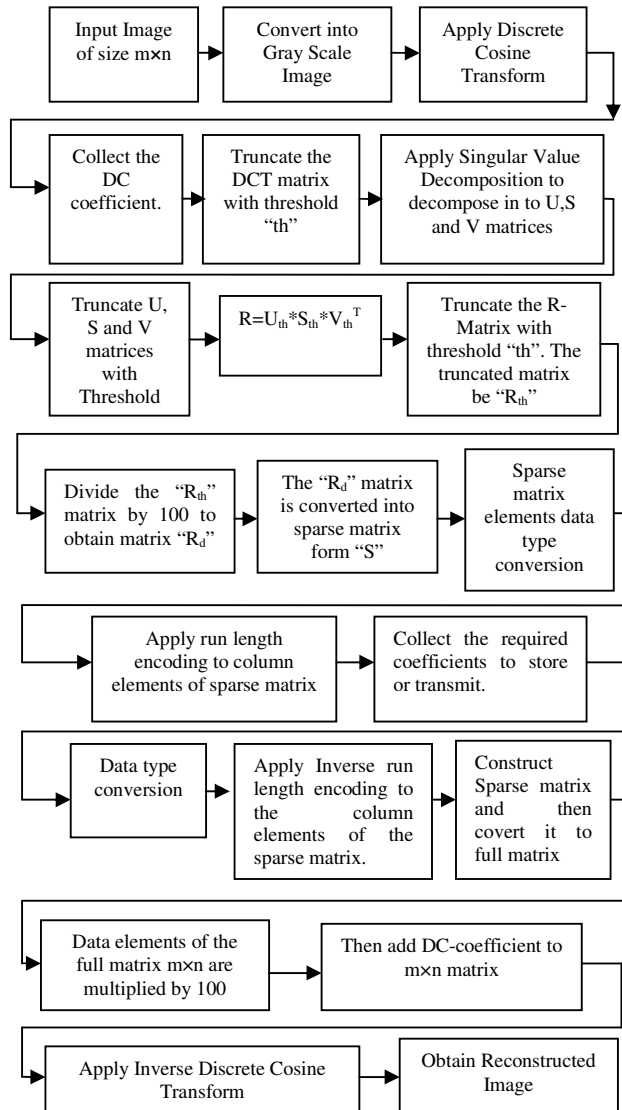




Figure1. Block diagram of the image compression using DCT-SVD-RLE method.

In this paper an effort is made to compress the image using hybrid compression techniques. They are (i) Image Compression using DCT-SVD-RLE method. (ii) Image compression using DCT-SVD method (iii) Image compression using DCT-RLE method, and (iv) Image compression using DCT method.

The figure1 shows the block diagram of the image compression using DCT-SVD-RLE method. In this method an image of size  $m \times n$  is applied as input. Suppose, if the given image is not in the Gray Scale format, it is converted into Gray Scale format. To this gray scale image, DCT is applied. Let this Discrete Cosine Transformed matrix be “D”. In the matrix “D”, the DC-coefficient is taken out and stored separately. The DC coefficient is the largest and important coefficient. Therefore it is stored separately. Then in the matrix “D”, all those coefficients less than the threshold “th” are neglected. After neglecting the coefficients less than “th” the discrete cosine transformed, truncated matrix “ $D_{th}$ ” is obtained. To this “ $D_{th}$ ” matrix, singular value decomposition is applied. This singular value decomposition, decomposes matrix “ $D_{th}$ ” into matrix “U”, matrix “S” and matrix “V”. Then, in the “U” matrix all those coefficients less than 0.02 are neglected. Let this truncated matrix be “ $U_{th}$ ”. In the “S” matrix all those coefficients less than 400 are neglected. Let this truncated matrix be “ $S_{th}$ ”. In the “V” matrix all those coefficients less than 0.05 are neglected. Let this truncated matrix be “ $V_{th}$ ”. The threshold value for “U” matrix i.e 0.02, the threshold value for “S” matrix i.e 400 and the threshold value for “V” matrix i.e 0.05 are selected empirically. The experiments are conducted for different values but these values found to be optimum. Then truncated matrices are multiplied such that  $R = U_{th} * S_{th} * V_{th}^T$ . Then “R” Matrix is again truncated with threshold “th”. That means all those coefficients less than “th” are neglected. Let this truncated matrix be “ $R_{th}$ ”. The elements of “ $R_{th}$ ” matrix are divided by 100 as a quantization. This results in matrix “ $R_d$ ”. The “ $R_d$ ” matrix contains most of the elements as zero, few coefficients are non-zero elements. Now “ $R_d$ ” matrix is converted as a sparse matrix. For example if

$$R_d = \begin{bmatrix} 0 & 91 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \text{ can be represented as}$$

S =

Row	Column	Data Element
1	2	91

Figure 2. Sparse Matrix and its representation.

The element under the column “Row” represents data element’s Row, the element under the column “Column” represents the data element’s column and the element under the column “Data element” represents element’s value.

This sparse notation reduces the memory required to store the data. Also, the sparse notation reduces the number of coefficients to be transmitted after image compression. The elements in the sparse matrix form are represented in double data type which requires 8-bytes for storage in Matlab. Therefore the elements under the column “Row” and “Column” in figure 2 are converted from double data type to int16 data type, where int16 data type requires 2-bytes for storage in Matlab. The elements present under the column “data element” are converted into “int8”, because int8 requires 1-byte to store the data in Matlab. Then for the elements present under the column “Column” in Figure.2 are run length encoded. Let the number of elements after run length encoding be “rlecol”. Therefore the number of bytes present in the compressed form is

$$dcomp = rowele * 2 + rlecol * 2 + Dataelemnts * 1 + dcf + r + c \quad (3)$$

Where dcomp=number of bytes in the compressed form. rowele=number of elements present in sparse notation under the column “Row”. rlecol= number of elements after run length encoding of elements under the column “column” in sparse matrix notation . Dataelemnts=number of elements under the column “Data element” in sparse matrix notation. dcf = 8 bytes to accommodate DC coefficient. r = 2bytes to accommodate the number of rows of the input image and c = 2bytes to accommodate the number of columns of the input image. Eventually “dcomp” is the number of bytes of input image in the compressed form. “rowele” and “rlecol” are multiplied by 2 because they are represented in the data type int16. It is assumed that “dcomp” is the number of bytes transmitted and the same number of coefficients are received. At the receiver, the received elements data types are converted back. The “rowele” , “rlecol” and



“Dataelemnts” are converted into double data type. Then inverse run-length encoding is applied to “rlecol” elements to get back the elements in sparse matrix notation. Then sparse matrix form is converted into full matrix form. Let this matrix be “RR” of size  $m \times n$ . Then each element of the matrix “RR” are multiplied by 100. Then DC coefficient is added. After this, Inverse Discrete cosine Transform is applied to reconstruct the image. The parameters such as MSE, PSNR and Compression ratio are evaluated.

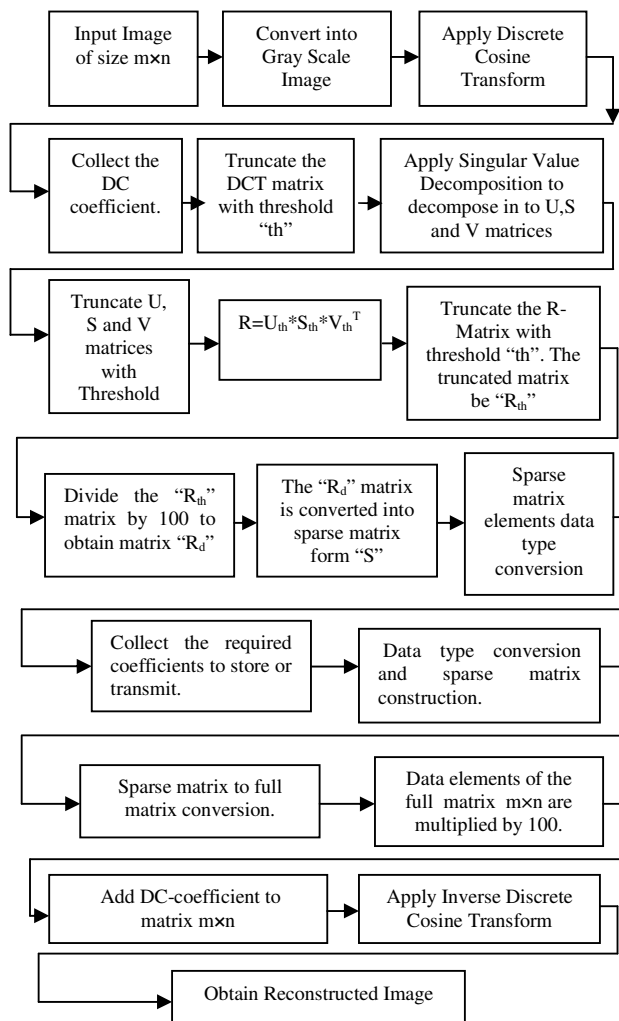


Figure3. Block diagram of the image compression using DCT-SVD method

The above figure shows the block diagram of the image compression using DCT-SVD method. In this method an image of size  $m \times n$  is applied as input. Suppose, if the given image is not in the Gray Scale format, it is converted into Gray Scale format. To

this gray scale image DCT is applied. Let this Discrete Cosine Transformed matrix be “D”. In the matrix “D”, the DC-coefficient is taken out and stored separately. Then in the matrix “D”, all those coefficients less than the threshold “th” are neglected. After neglecting the coefficients less than “th”, the discrete cosine transformed, truncated matrix “D<sub>th</sub>” is obtained. To this “D<sub>th</sub>” matrix singular value decomposition is applied. This singular value decomposition, decomposes matrix “D<sub>th</sub>” into matrix “U”, matrix “S” and matrix “V”. Then, in the “U” matrix all those coefficients less than  $th_u = 0.02$  are neglected. Let this truncated matrix be “U<sub>th</sub>”. In the “S” matrix all those coefficients less than  $th_s = 400$  are neglected. Let this truncated matrix be “S<sub>th</sub>”. In the “V” matrix all those coefficients less than  $th_v = 0.05$  are neglected. Let this truncated matrix be “V<sub>th</sub>”. The threshold value for “U” matrix i.e 0.02, the threshold value for “S” matrix i.e 400 and the threshold value for “V” matrix i.e 0.05 are selected empirically. The experiments are conducted for different values but these values found to be optimum. Then truncated matrices are multiplied such that  $R = U_{th} * S_{th} * V_{th}^T$ . Then “R” Matrix is again truncated with threshold “th”. That means all those coefficients less than “th” are neglected. Let this truncated matrix be “R<sub>th</sub>”. The elements of “R<sub>th</sub>” matrix are divided by 100 as a quantization. This results in matrix “R<sub>d</sub>”. The “R<sub>d</sub>” matrix contains most of the elements as zero, few coefficients are non-zero elements. Now “R<sub>d</sub>” matrix is converted as a sparse matrix. Then the elements under the “Row” and the “Column” in figure 2 are converted from double data type to int16 data type. The elements present under the column “data element” are converted into “int8”. Therefore the number of bytes present in the compressed form is

$$dcomp = rowele * 2 + colele * 2 + Dataelemnts * 1 + dcf + r + c \quad (4)$$

dcomp=number of bytes in the compressed form. rowele =number of elements present in the sparse notation under the column “Row”. colele =number of elements present in the sparse notation under the column “Column”. Dataelemnts=number of elements present under the column “Data element” in sparse matrix notation. dcf = 8 bytes to accommodate the DC coefficient. r = 2bytes to accommodate the number of rows of the input image and c = 2bytes to accommodate the number of columns of the input image. “rowele” and “colele”

are multiplied by 2 because they are represented in the data type int16. It is assumed that “dcomp” is the number of bytes transmitted and the same number of coefficients are received. At the receiver, the received elements data types are converted back. The “rowele”, “colele” and “Dataelemnts” are converted into double data type. Then sparse matrix is constructed. Then sparse matrix form is converted into full matrix form. Let this matrix be “RR” of size  $m \times n$ . Then each element of the matrix “RR” are multiplied by 100. Then DC coefficient is added. After this, Inverse Discrete cosine Transform is applied to reconstruct the image. The parameters such as MSE, PSNR and Compression ratio are evaluated.

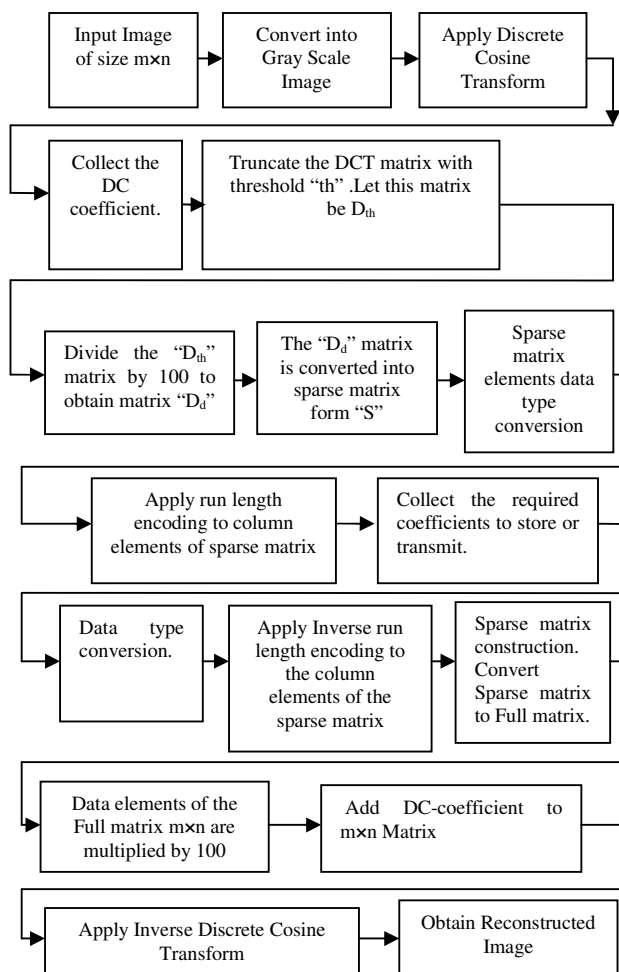


Figure4. Block diagram of the image compression using DCT-RLE method

The above figure shows the block diagram of the image compression using DCT-RLE method. In this method an image of size  $m \times n$  is applied as input.

Suppose, if the given image is not in the Gray Scale format, it is converted into Gray Scale format. To this gray scale image DCT is applied. Let this Discrete Cosine Transformed matrix be “D”. In the matrix “D”, the DC-coefficient is taken out and stored separately. Then in the matrix “D”, all those coefficients less than the threshold “th” are neglected. After neglecting the coefficients less than “th” in the discrete cosine transformed, truncated matrix “D<sub>th</sub>” is obtained. The elements of “D<sub>th</sub>” matrix are divided by 100 as a quantization. This results in matrix “D<sub>d</sub>”. The “D<sub>d</sub>” matrix contains most of the elements as zero, few coefficients are non-zero elements. Now “D<sub>d</sub>” matrix is converted as a sparse matrix. The elements present under the column “data element” of sparse notation are converted into “int8”, because int8 requires 1-byte to store the data in Matlab. The elements present under the column “Row” and “Column” of sparse notation are converted from data type double to int16. Where the data type “int16” requires 2-bytes. Then for the elements present under the column “column” in Figure.2 are run length encoded. Let the number of elements after run length encoding be “rlecol”. Therefore the number of bytes present in the compressed form is

$$dcomp = rowele * 2 + rlecol * 2 + Dataelemnts * 1 + dcf + r + c \quad (5)$$

Where dcomp=number of bytes in the compressed form. rowele=number of elements present in sparse notation under the column “Row”. rlecol= number of elements after run length encoding of elements under the column “column” in sparse matrix notation. Dataelemnts=number of elements under the column “Data element” in sparse matrix notation. dcf= 8 bytes to accommodate DC coefficient. r = 2bytes to accommodate number of rows of the input image and c = 2 bytes to accommodate number of columns of the input image. Eventually “dcomp” is the number of bytes of input image in the compressed form. “rowele” and “rlecol” are multiplied by 2 because they are represented in the data type int16. It is assumed that “dcomp” is the number of bytes transmitted and it is assumed that same number of coefficients are received. At the receiver, the received elements data types are converted back. The “rowele”, “rlecol” and “Dataelemnts” are converted into double data type. Then inverse run-length encoding is applied to “rlecol” elements to get back the elements in sparse

matrix notation. Then sparse matrix form is converted into full matrix form. Let this matrix be “RR” of size  $m \times n$ . Then each element of the matrix “RR” are multiplied by 100. Then DC coefficient is added. After this, Inverse Discrete cosine Transform is applied to reconstruct the image. The parameters such as MSE, PSNR and Compression ratio are evaluated.

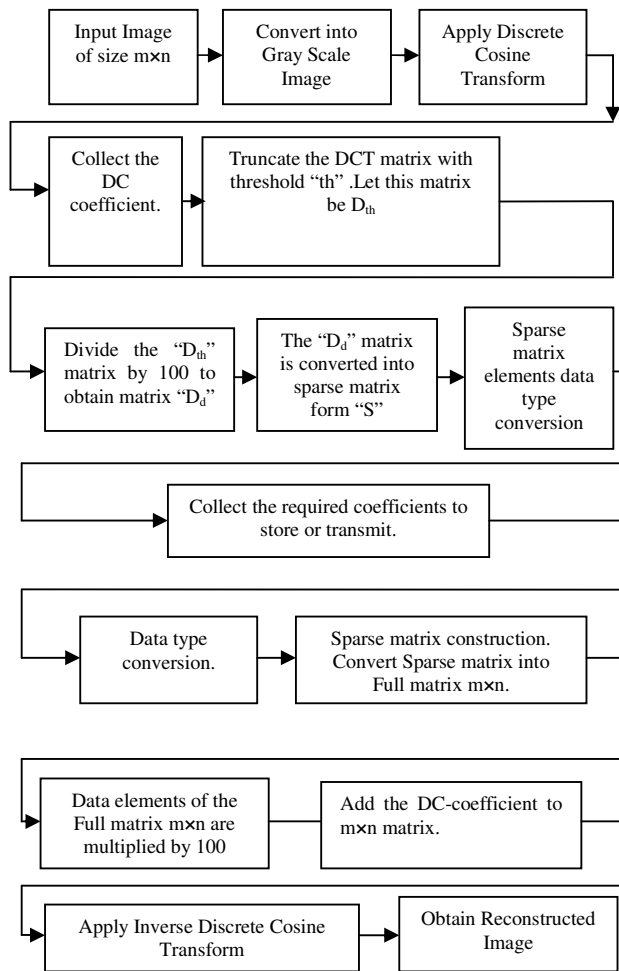


Figure5. Block diagram of the image compression using DCT method

The above figure shows the block diagram of the image compression using DCT method. In this method an image of size  $m \times n$  is applied as input. Suppose, if the given image is not in the Gray Scale format, it is converted into Gray Scale format. To this gray scale image DCT is applied. Let this Discrete Cosine Transformed matrix be “D”. In the matrix “D”, the DC-coefficient is taken out and stored separately. Then in the matrix “D”, all those coefficients less than the threshold “th” are neglected. After neglecting the coefficients less than

“th”, discrete cosine transformed, truncated matrix “D<sub>th</sub>” is obtained. The elements of “D<sub>th</sub>” matrix are divided by 100 as a quantization. This results in matrix “D<sub>d</sub>”. The “D<sub>d</sub>” matrix contains most of the elements as zero, few coefficients are non-zero elements. Now “D<sub>d</sub>” matrix is converted as a sparse matrix. Therefore the elements under the “Row” and the “Column” in sparse matrix notation are converted from double data type to int16 data type. The elements present under column “data element” are converted into “int8”. Therefore the number of bytes present in the compressed form is

$$dcomp = rowele * 2 + colele * 2 + Dataelemnts * 1 + dcf + r + c \quad (6)$$

dcomp=number of bytes in the compressed form. rowele =number of elements present in sparse matrix notation under the column “Row”. colele =number of elements present in sparse matrix notation under the column “Column”. Dataelemnts = number of elements under the column “Data element” in sparse matrix notation. dcf = 8 bytes to accommodate DC coefficient. r = 2bytes to accommodate the number of rows of the input image and c = 2bytes to accommodate the number of columns of the input image. “rowele” and “colele” are multiplied by 2 because they represented in the data type int16. It is assumed that “dcomp” number of bytes are transmitted and the same number of coefficients are received. At the receiver, the received elements data types are converted. The “rowele”, “colele” and “Dataelemnts” are converted into double data type. Then sparse matrix is constructed. Then sparse matrix form is converted into full matrix form. Let this matrix be “RR” of size  $m \times n$ . Then each element of the matrix “RR” are multiplied by 100. Then DC coefficient is added. After this, Inverse Discrete Cosine Transform is applied to reconstruct the image. The parameters such as MSE, PSNR and Compression ratio are evaluated.

#### IV. IMPLEMENTATION

The experimentation is conducted using Matlab 7.6 on Intel(R) core i3 processor at 2.4 GHz. There are four methods. They are (i) Image Compression using DCT-SVD-RLE method. (ii) Image compression using DCT-SVD method (iii) Image compression using DCT-RLE method, and (iv) Image compression using DCT method.

The algorithm of the Image Compression using DCT-SVD-RLE method is as follows.

- (i) Accept an image of size “m” rows and “n” columns.
  - (ii) If the accepted image is in any form other than the Gray Scale format, convert it into Gray Scale format.
  - (iii) Apply Discrete Cosine Transform to the Gray Scale format. This results in the matrix “D”.
  - (iv) Collect the DC coefficient of the Discrete Cosine transformed matrix separately.
  - (v) Apply threshold “th” to the Discrete Cosine Transformed matrix “D”.i.e all those coefficients less than “th” in the “D” matrix are neglected. This results in the matrix “D<sub>th</sub>”.
  - (vi) Apply Singular Value Decomposition to the matrix “D<sub>th</sub>”. This results in three matrices “U”, “S” and “V”.
  - (vii) Apply threshold “th<sub>u</sub> = 0.02” to the “U” matrix. i.e all those coefficients less than “th<sub>u</sub>” are neglected. This results in the matrix “U<sub>th</sub>”. Apply threshold “th<sub>s</sub>=400” to the “S” matrix. i.e all those coefficients less than “th<sub>s</sub>” are neglected. This results in the matrix “S<sub>th</sub>”. Similarly, apply threshold “th<sub>v</sub> =0.05” to the “V” matrix. i.e all those coefficients less than “th<sub>v</sub>” are neglected. This results in the matrix “V<sub>th</sub>”.
  - (viii) Then multiply the matrices such that  $R = U_{th} * S_{th} * V_{th}^T$ .
  - (ix) Then apply threshold “th” to the matrix “R”.i.e all those coefficient less than “th” are neglected in this matrix “R”. This results in the matrix “R<sub>th</sub>”.
  - (x) Divide every element of the matrix “R<sub>th</sub>” by 100.This results in the matrix “R<sub>d</sub>”.
  - (xi) Convert Matrix “R<sub>d</sub>” from full matrix form to sparse matrix form. Let this sparse matrix be “S”
  - (xii) In the sparse matrix form “S”, convert the data elements under the column “Row” from data type double to the int16, convert the data elements under the column “column” from data type double to the int16 and convert the data elements under the column “data element” from data type double to the int8.
  - (xiii) After this, apply run length encoding to the data under the column “column” of the sparse matrix. This gives the run length encoded data as “rlecol”
  - (xiv) Then, the elements of the column “data elements” of the sparse matrix, the elements of the column “row” of the sparse matrix and the run length encoded data “rlecol”, DC-coefficient ,number of rows of matrix “R<sub>d</sub>” and the number of rows and columns of the matrix “R<sub>d</sub>” are transmitted.
  - (xv) At the receiver, it is assumed that all the coefficients are received. Convert the elements under the column “row” from data type “int16” to double, convert the elements of the run length encoded “rlecol” data from data type “int16” to double and convert the elements under the column “data element” from data type “int8” to double. Then, inverse run length encoding is applied to the elements “rlecol” to obtain the elements under the column “column” of the sparse matrix.
  - (xvi) Then construct the sparse matrix.
  - (xvii) Then convert the sparse matrix into full matrix. Let this matrix be “RR”.
  - (xviii) Multiply each element of “RR” by 100.Then add the DC-coefficient to the matrix “RR”.
  - (xix) Apply inverse discrete cosine transform to the matrix “RR” to obtain the reconstructed image.
  - (xx) Then parameters such as MSE, PSNR and Compression Ratio are evaluated. The mathematical equations for MSE,PSNR and compression ratio are as follows,
- $$MSE = \frac{\sum_{i=1}^m \sum_{j=1}^n |a(i,j) - b(i,j)|^2}{m \times n} \quad (7)$$
- where m=number of rows of the image, n= number of columns of the image, a(i,j)= The element of the original image matrix at the i<sup>th</sup> row and j<sup>th</sup> column, b(i,j) is The element of the reconstructed image matrix at the i<sup>th</sup> row and j<sup>th</sup> column.
- The Peak Signal to Noise Ratio is given by
- $$PSNR = 10 \log_{10} \frac{255^2}{MSE} \quad (8)$$
- Where MSE=Mean Squared Error
- The compression ratio is given by
- $$CR = \frac{m \times n}{dcomp} \quad (9)$$
- Where
- $$dcomp = rowele * 2 + rlecol * 2 + Dataelemnts * 1 + dcf + r + c \quad (10)$$

dcomp=number of bytes in the compressed form.  
 rowele =number of elements present in sparse matrix notation under the column "Row". rlecol= number of elements after run length encoding of elements under the column "column" in Sparse matrix notation. Dataelemnts = number of elements present under the column "Data element" in sparse matrix notation.dcf= 8 bytes to accommodate DC coefficient. r= 2bytes to accommodate the number of rows of the input image and c= 2bytes to accommodate the number of columns of the input image. m= number of rows of the input image matrix, n=number of columns of the input image matrix.

(xxi) Then different value of "th" is set and steps from (v) to (xx) are repeated.

The algorithm for the image compression using DCT-SVD method is as follows.

- (i) Accept an image of size "m" rows and "n" columns.
- (ii) If the accepted image is in any form other than the Gray Scale format convert it into Gray Scale format.
- (iii) Apply Discrete Cosine Transform to the Gray Scale format. This results in the matrix "D".
- (iv) Collect the DC coefficient of the Discrete Cosine transformed matrix separately.
- (v) Apply threshold "th" to the Discrete Cosine Transformed matrix "D". i.e all those coefficients less than "th" in the "D" matrix are neglected. This results in the matrix "D<sub>th</sub>".
- (vi) Apply Singular Value Decomposition to the matrix "D<sub>th</sub>". This results in three matrices "U", "S" and "V".
- (vii) Apply threshold "th<sub>u</sub>" = 0.02 to the "U" matrix. i.e all those coefficients less than "th<sub>u</sub>" are neglected. This results in the matrix "U<sub>th</sub>". Apply threshold "th<sub>s</sub>=400" to the "S" matrix. i.e all those coefficients less than "th<sub>s</sub>" are neglected. This results in the matrix "S<sub>th</sub>". Similarly, apply threshold "th<sub>v</sub>=0.05" to the "V" matrix. i.e all those coefficients less than "th<sub>v</sub>" are neglected. This results in the matrix "V<sub>th</sub>".
- (viii) Then multiply the matrices such that  $R = U_{th} * S_{th} * V_{th}^T$ .

(ix) Then apply threshold "th" to the matrix "R".i.e all those coefficient less than "th" are neglected in this matrix "R". This results in the matrix "R<sub>th</sub>".

(x) Divide every element of the matrix "R<sub>th</sub>" by 100.This results in the matrix "R<sub>d</sub>".

(xi) Convert Matrix "R<sub>d</sub>" from full matrix form to sparse matrix form. Let this sparse matrix be "S".

(xii) In the sparse matrix form "S", convert the data elements under the column "Row" from data type double to the int16, convert the data elements under the column "column" from data type double to the int16 and convert the data elements under the column "data element" of sparse matrix notation from data type double to the int8.

(xiii) Then, the elements of the column "data elements" of the sparse matrix, the elements of the column "row" of the sparse matrix and the elements of the column "column" of the sparse matrix, DC-coefficient and the number of rows and columns of the Input image matrix are transmitted.

(xiv) At the receiver, it is assumed that all the coefficients are received.

(xv) Convert the elements under the column "row" of the sparse matrix from data type "int16" to double, convert the elements under the column "column" of the sparse matrix from data type "int16" to double and convert the elements under the column "data element" from data type "int8" to double.

(xvi) Then construct the sparse matrix.

(xvii) Then convert the sparse matrix into full matrix. Let this matrix be "RR".

(xviii) Multiply each element of "RR" by 100.Then, add the DC-coefficient to the matrix "RR".

(xix) Apply inverse discrete cosine transform to the matrix "RR" to obtain the reconstructed image.

(xx) Then parameters such as MSE, PSNR and Compression Ratio are evaluated. MSE and PSNR are calculated with formula as explained above, where as compression ratio is calculated by

$$CR = \frac{m \times n}{dcomp} \quad (11)$$

Where

$dcomp = rowele * 2 + colele * 2 + Dataelemnts * 1 + dcf + r + c$  (12)  
 dcomp=number of bytes in the compressed form.  
 rowele =number of elements present in sparse matrix notation under the column "Row".  
 colele =number of elements present in sparse matrix notation under the column "Column".  
 Dataelemnts=number of elements present under the



column "Data element" in sparse matrix notation. dcf = 8 bytes to accommodate DC coefficient. r = 2bytes to accommodate the number of rows of the input image and c = 2bytes to accommodate the number of columns of the input image. m= number of rows of the input image matrix, n=number of columns of the input image matrix.

(xxi) Then different value of "th" is set and steps from (v) to (xx) are repeated.

The algorithm of the Image compression using DCT-RLE method is as follows.

- (i) Accept an image of size "m" rows and "n" columns.
- (ii) If the accepted image is in any form other than the Gray Scale format convert it into Gray Scale format.
- (iii) Apply Discrete Cosine Transform to the Gray Scale format. This results in the matrix "D".
- (iv) Collect the DC coefficient of the Discrete Cosine transformed matrix separately.
- (v) Apply threshold "th" to the Discrete Cosine Transformed matrix "D". i.e all those coefficients less than "th" in the "D" matrix are neglected. This results in the matrix "D<sub>th</sub>".
- (vi) Divide every element of the matrix "D<sub>th</sub>" by 100. This results in the matrix "D<sub>d</sub>".
- (vii) Convert Matrix "D<sub>d</sub>" from full matrix form to sparse matrix form. Let this sparse matrix be "S"
- (viii) In the sparse matrix form "S", convert the data elements under the column "Row" from data type double to the int16, convert the data elements under the column "column" from data type double to the int16 and convert the data elements under column "data element" from data type double to the int8.
- (ix) After this, apply run length encoding to the data under the column "column" of the sparse matrix. This gives the run length encoded data as "rlecol".
- (x) Then, the elements of the column "data elements" of the sparse matrix, the elements of the column "row" of the sparse matrix and the run length encoded data "rlecol", DC-coefficient, number of rows of Input Image matrix and the

number of columns of the Input Image matrix are transmitted.

At the receiver, it is assumed that all the coefficients are received. Convert the elements under column "row" from data type "int16" to double, convert the element "rlecol" from data type "int16" to double and convert the elements under the column "data element" from data type "int8" to double. First inverse run length encoding is applied to elements "rlecol" to obtain the elements under the column "column" of the sparse matrix.

(xi) Then construct the sparse matrix.

(xii) Then convert the sparse matrix into full matrix. Let this matrix be "RR".

(xiii) Multiply each element of "RR" by 100. Then add the DC-coefficient to the matrix "RR".

(xiv) Apply inverse discrete cosine transform to the matrix "RR" to obtain the reconstructed image.

(xv) Then parameters MSE, PSNR and Compression ratio are evaluated by equation (7), (8) and (9) respectively.

(xvi) Then different value of "th" is set and steps from (v) to (xv) are repeated.

Algorithm of the Image Compression using DCT

method is as follows.

- (i) Accept an image of size "m" rows and "n" columns.
- (ii) If the accepted image is in any form other than the Gray Scale format, convert it into Gray Scale format.
- (iii) Apply Discrete Cosine Transform to the Gray Scale format. This results in the matrix "D".
- (iv) Collect the DC coefficient of the Discrete Cosine transformed matrix separately.
- (v) Apply threshold "th" to the Discrete Cosine Transformed matrix "D". i.e all those coefficients less than "th" in the "D" matrix are neglected. This results in the matrix "D<sub>th</sub>".
- (vi) Divide every element of the matrix "D<sub>th</sub>" by 100. This results in the matrix "D<sub>d</sub>".

- (vii) Convert Matrix “D<sub>d</sub>” from full matrix form to sparse matrix form. Let this sparse matrix be “S”
- (viii) In the sparse matrix form “S”, convert the data elements under the column “Row” from data type double to the int16, convert the data elements under the column “column” from data type double to the int16 and convert the data elements under column “data element” from data type double to the int8.
- (ix) Then, the elements of the column “data elements” of the sparse matrix, the elements of the column “row ” of the sparse matrix and the elements of the column “column” of the sparse matrix, DC-coefficient ,number of rows of Input image matrix and the number of rows and columns of the Input image matrix are transmitted.
- (x) At the receiver, it is assumed that all the coefficients are received.
- (xi) Convert the elements under the column “row” of the sparse matrix from data type “int16” to double, convert the elements under the column “column” of the sparse matrix from data type “int16” to double and convert the elements under the column “data element” from data type “int8” to double.
- (xii) Then construct the sparse matrix.
- (xiii) Then convert the sparse matrix into full matrix. Let this matrix be “RR”.
- (xiv) Multiply each element of “RR” by 100.Then, add the DC-coefficient to the matrix “RR”.
- (xv) Apply inverse discrete cosine transform to the matrix “RR” to obtain the reconstructed image.
- (xvi) Then parameters such as MSE, PSNR and Compression Ratio are evaluated using equations (7), (8) and (11).
- (xvii) Then different value of “th” is set and steps from (v) to (xvi) are repeated.

## V. RESULTS AND DISCUSSIONS

Experiments are conducted for different set of inputs by considering different resolution and

different file formats such as .tiff, .png, .jpg etc. A sample of the experimental result is displayed for further discussion and analysis.

The details of the input image and its results are as follows.

Image name: river.jpg

Image size: 425x318

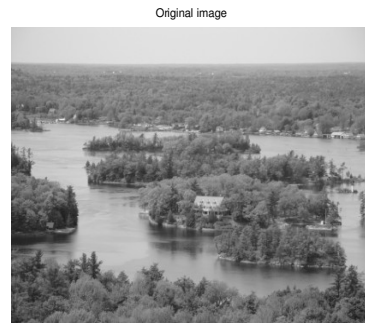


Figure6. Input Image.

Different trails of experimentation in each method are tabulated as follows.

TABLE 1.RESULTS OF THE IMAGE COMPRESSION USING DCT-SVD-RLE METHOD.

th<sub>u</sub>=0.02, th<sub>s</sub>=400, th<sub>v</sub>=0.05

Sl. No.	th	MSE	PSNR (in dB)	Memory (in Bytes)	CR
1	50	187.9812	25.3897	4833	27.9640
2	52	190.5455	25.3308	4532	29.8213
3	54	190.8043	25.3249	4407	30.6671
4	56	195.6602	25.2158	4095	33.0037
5	58	195.5874	25.2174	3948	34.2325

The above table shows the results of the image compression using DCT-SVD-RLE method. “th” is the threshold applied to truncate the Discrete Cosine Transform matrix, “MSE” is the Mean Squared Error. “PSNR” is the Peak Signal to Noise Ratio. “Memory” is the number of bytes of the image in the compressed form. “CR” is the Compression Ratio. “ th<sub>u</sub> ” is the threshold applied for the “U”-Matrix , “th<sub>s</sub>” is the threshold applied for the “S”-Matrix , “th<sub>v</sub>” is the threshold applied for the “V”-Matrix.

TABLE 2.RESULTS OF THE IMAGE COMPRESSION USING DCT-SVD METHOD. th<sub>u</sub>=0.02, th<sub>s</sub>=400,th<sub>v</sub>=0.05

Sl. No.	th	MSE	PSNR (in dB)	Memory (in Bytes)	CR
1	50	187.9812	25.3897	7547	17.9078
2	52	190.5455	25.3308	7072	19.1106
3	54	190.8043	25.3249	6857	19.7098
4	56	195.6602	25.2158	6357	21.2600
5	58	195.5874	25.2174	6112	22.1122

The above table shows the results of the image compression using DCT-SVD method. “th” is the threshold applied to truncate the Discrete Cosine Transform matrix, “MSE” is the Mean Squared Error. “PSNR” is the Peak Signal to Noise Ratio. “Memory” is the number of bytes of the image in the compressed form. “CR” is the Compression Ratio. “ $th_u$ ” is the threshold applied for the “U”-Matrix, “ $th_s$ ” is the threshold applied for the “S”-Matrix, “ $th_v$ ” is the threshold applied for the “V”-Matrix.

TABLE 3. RESULTS OF THE IMAGE COMPRESSION USING DCT-RLE METHOD.

Sl. No.	th	MSE	PSNR (in dB)	Memory (in Bytes)	CR
1	50	121.9080	27.2705	8784	15.3859
2	52	122.1413	27.2622	8290	16.3028
3	54	122.7838	27.2394	7846	17.2253
4	56	123.8261	27.2027	7395	18.2759
5	58	125.0919	27.1585	7005	19.2934

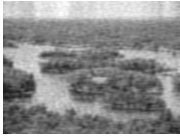
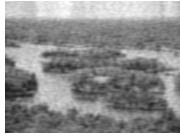
The above table shows the results of the image compression using DCT-RLE method. “th” is the threshold applied to truncate the Discrete Cosine Transform matrix, “MSE” is the Mean Squared Error. “PSNR” is the Peak Signal to Noise Ratio. “Memory” is the number of bytes of the image in the compressed form. “CR” is the Compression Ratio.

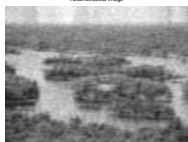
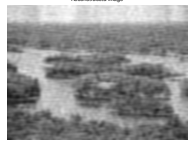
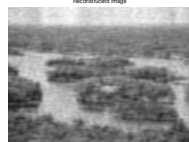
TABLE 4. RESULTS OF THE IMAGE COMPRESSION USING DCT METHOD.

Sl. No.	th	MSE	PSNR (in dB)	Memory (in Bytes)	CR
1	50	121.9080	27.2705	13692	9.8707
2	52	122.1413	27.2622	12902	10.4751
3	54	122.7838	27.2394	12182	11.0942
4	56	123.8261	27.2027	11457	11.7963
5	58	125.0919	27.1585	10847	12.4597

The above table shows the results of the image compression using DCT method. “th” is the threshold applied to truncate the Discrete Cosine Transform matrix, “MSE” is the Mean Squared Error. “PSNR” is the Peak Signal to Noise Ratio. “Memory” is the number of bytes of the image in the compressed form. “CR” is the Compression Ratio.

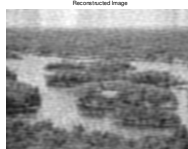
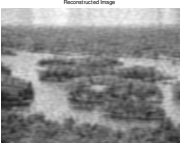
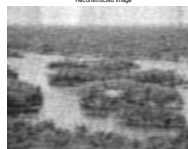
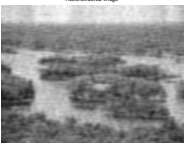
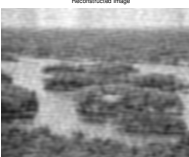
TABLE 5 RECONSTRUCTED IMAGES OF THE IMAGE COMPRESSION USING DCT-SVD-RLE METHOD.

 th=50, PSNR=25.3897,	 th=52, PSNR=25.3308,
---	---

CR=27.9640	CR=29.8213
 th=54, PSNR=25.3249, CR=30.6671	 th=56, PSNR=25.2158, CR=33.0037
 th=58, PSNR=25.2174, CR=34.2325	

The above table shows the reconstructed images of the Image Compression using DCT-SVD-RLE method. “th” is the threshold, PSNR= Peak Signal to Noise Ratio and CR=Compression ratio.






TABLE 6 RECONSTRUCTED IMAGES OF THE IMAGE COMPRESSION USING DCT-SVD METHOD.

 th=50, PSNR=25.3897, CR=17.9078	 th=52, PSNR=25.3308, CR=19.1106
 th=54, PSNR=25.3249, CR=19.7098	 th=56, PSNR=25.2158, CR=21.2600
 th=58, PSNR=25.2174, CR=22.1122	

The above table shows the reconstructed images of the Image Compression using DCT-SVD method.





“th” is the threshold, PSNR= Peak Signal to Noise Ratio and CR=Compression ratio.

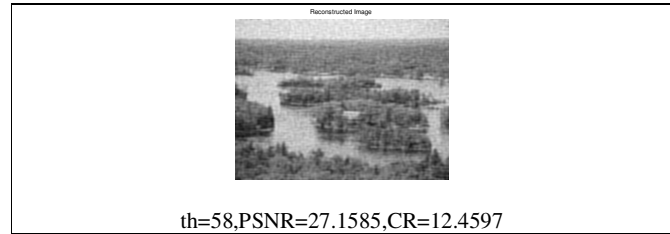
TABLE7 . RECONSTRUCTED IMAGES OF THE IMAGE COMPRESSION USING DCT-RLE METHOD.

 <p>th=50,PSNR=27.2705, CR=15.3859</p>	 <p>th=52,PSNR=27.2622, CR=16.3028</p>
 <p>th=54,PSNR=27.2394, CR=17.2253</p>	 <p>th=56,PSNR=27.2027, CR=18.2759</p>
 <p>th=58,PSNR=27.1585,CR=19.2934</p>	

The above table Shows the reconstructed images of the Image Compression using DCT-RLE method. “th” is the threshold, PSNR= Peak Signal to Noise Ratio and CR=Compression ratio.

TABLE 8 . RECONSTRUCTED IMAGES OF THE IMAGE COMPRESSION USING DCT METHOD.

 <p>th=50,PSNR=27.2705,CR=9.870 7</p>	 <p>th=52,PSNR=27.2622,CR=10.47 51</p>
 <p>th=54,PSNR=27.2394,CR=11.09 42</p>	 <p>th=56,PSNR=27.2027,CR=11.79 63</p>



The above table shows the reconstructed images of the Image Compression using DCT method. “th” is the threshold, PSNR= Peak Signal to Noise Ratio and CR=Compression ratio.

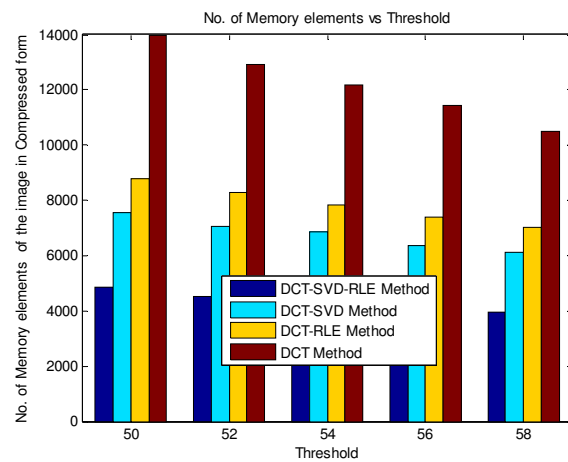


Figure7. Graph of No. Of memory elements Vs Threshold

The above figure shows the graph of No. of memory elements of the image in the compressed form vs Threshold. It can be observed that for the given threshold, the number of memory bytes to be stored in the compressed form or to be transmitted in the DCT-SVD-RLE method is very less compared to other three methods. As the threshold increases number of memory elements to be stored or transmitted decreases. It can be seen that for a threshold 58 the number of memory elements of compressed image is 3948 bytes in DCT-SVD-RLE method, but for the same threshold the number of memory elements of the compressed image are 6112 bytes, 7005 bytes and 10847 bytes in DCT-SVD method, DCT-RLE method and DCT method respectively. Therefore DCT-SVD-RLE method is more efficient.

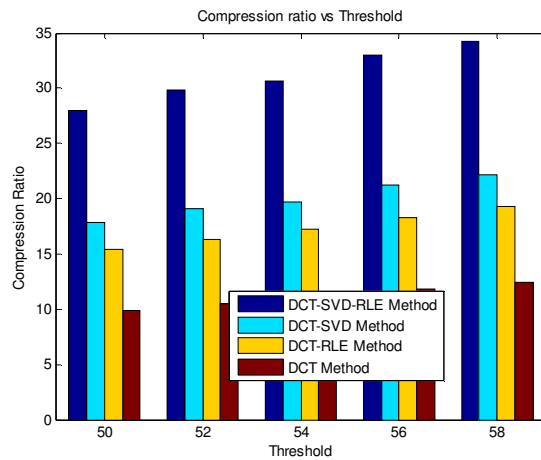


Figure 8. Graph of Compression ratio Vs Threshold

The above figure shows the graph of Compression ratio vs Threshold. In the above figure for the given threshold compression ratio in the DCT-SVD-RLE method is more compared to other three methods. As the threshold increases the compression ratio also increases. It can be seen that for a threshold 58 compression ratio is 34.2325 in DCT-SVD-RLE method, but for the same threshold compression ratio are 22.1122, 19.2934 and 12.4597 in DCT-SVD method, DCT-RLE method and DCT method respectively. Here also results shows that DCT-SVD-RLE method is more efficient

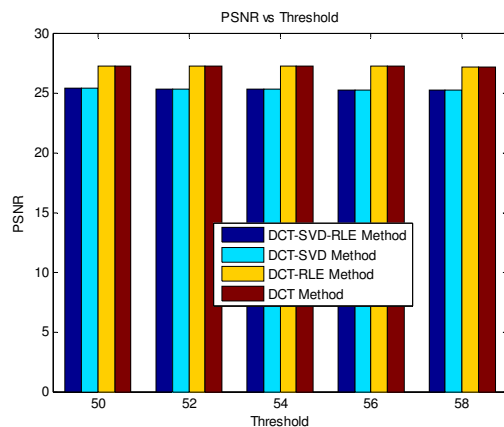


Figure 9. Graph of PSNR Vs Threshold

The above figure shows the graph of PSNR vs Threshold. In the above figure it can be seen that as the threshold increase PSNR decreases. PSNR of the DCT-SVD-RLE method is lesser than DCT-RLE method. Since we are able to achieve a good compression Ratio with DCT-SVD-RLE method, the price paid happens to be PSNR.

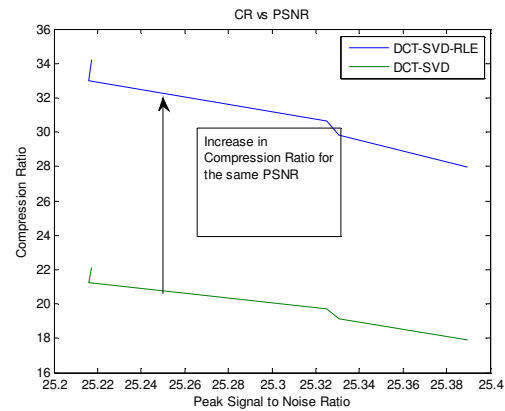


Figure 10. Graph of Compression ratio Vs PSNR

In the above figure, graph of Compression ratio Vs PSNR is shown. For the same PSNR, Compression Ratio in DCT-SVD-RLE method is more than the DCT-SVD method. Therefore DCT-SVD-RLE method is good.

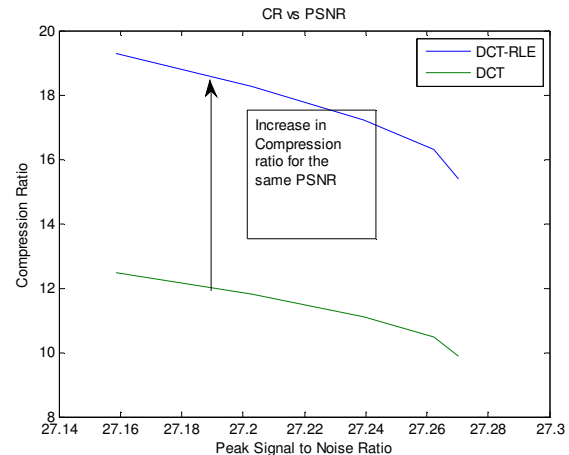


Figure 11. Graph of Compression ratio Vs PSNR

In the above figure, graph of Compression ratio vs PSNR is shown. For the same PSNR, Compression Ratio in DCT-RLE method is more than the DCT method.

But the compression ratio of DCT-RLE method is not so good as DCT-SVD-RLE method.

With all this, it can be shown that using DCT-SVD-RLE method good compression ratio can be obtained by losing PSNR around 2dB.



## VI. SCOPE FOR FURTHER ENHANCEMENT

Four methods are compared to achieve a good compression ratio for the fixed threshold. However the compression ratio can also be explored by using other transforms such as wavelet, KLT, Hadamard and slant. The experimentation can be extended by considering different combinations such as DCT-Hadamard, DCT-Wavelets, and DCT-slant.

## REFERENCES:

- [1] Raghavendra.M.J,Prasanth.H.S and S.Sandya, "DCT SVD Based Hybrid Transform Coding for Image Compression", International Journal of Recent and Innovative Trends in computing and communication. 2015
- [2] Prasanth.H.S,Shashidhara.H.L and Balasubramanyamurthy.K.N, "Image compression using SVD" Intenational Conference on Computational Intelligence and Multimedia applications, Vol.3,2007
- [3] S.Sridhar,P.Rajeshkumar and K.V.Ramanaiah , " Wavelet Transform Techniques for Image Compression-An evaluation " International Journal of Image,Graphics and Signal Processing,2014
- [4] T.D.Khadatre, Mayuri Chaudari, Sushma B, and Yogita Raut, "A combined novel approach for Image Compression using Vector Quantization and wavelet transform" International Journal of Application or Innovation in Engineering & Management.Vol.3, April 2014
- [5] Athira.M.S. and V.Kalaichelvi, "An Intelligent Technique for Image compression " International Journal of Recent Developments in Engineering and technology ,June 2014
- [6] Pallavi.M.Sune and Vijay.K.Shandilya, "Image Compression Techniques based on Wavelet and Huffman coding " International Journal of Advanced Research in Computer Science and Software Engineering, April 2013
- [7] E.Praveenkumar and M.G.sumithra, "Medical Image Compression using Integer Multi wavelets Transform for Telemedicine Applications " International Journal of Engineering and Computer Science ,May 2013
- [8] D.Vishnuvardhan, Sreenivasan.B and I.Suneetha. "Advanced Digital Image compression Technique using curvelet Transform" International Journal of Engineering Research and Applications Vol.3,Issue-4,Aug 2013
- [9] Birendrakumar Patel, Suyesh Agrawal, "Image Compression Techniques using artificial neural networks" International Journal of Advanced Research in Computer Engineering & Technology, Vol.2, October 2013
- [10] Sumegha yadav, Tarun kumar.R. "Transform Based Hybrid Image Compression Techniques in conjunction with Fractal Image compression scheme" International Journal of Advancements in Research & Technology, Volume 1, Issue 4 Aril 2013.
- [11] Rowayda A.Sadek , "SVD Based Image Processing Applications:State of the Art,Contributions and Research challenges", International Journal of Advanced Computer Science and Applications.Vol 3,2012
- [12] K.R.Rao ,Ahmed.N,Natarajan.T , "Discrete Cosine Transform", IEEE Transaction on Computers,1974

## AUTHOR'S PROFILE



Raghavendra.M.J obtained his Bachelor degree from Mysore University and Master Degree from NITK,Suratkal.His reserch interest includes Multimedia and Signal Processing.He is persuing research program in V.T.U.He is currently working as an Assistant professor in the Department of Telecommunication Engineering,PES Institute of Technology, Bangalore.



Dr.Prasantha.H.S received Bachelor degree from Bangalore University, Master Degree from V.T.U, Belgaum, and Ph.D from Anna University, Chennai, in the area of Multimedia and Image Processing. He has 16+ years of teaching and research experience. His research interest includes Multimedia and Signal Processing. He is currently guiding students for their research program in V.T.U and other university. Currently, he is working as a Professor in the department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore.



Dr.S.Sandya obtained her Ph.D from Indian Institute of Science, Bangalore. She has vast experience in the field of industry, research and teaching. Her research interest includes Satellite communications, Wireless Sensor Networks and Embedded Systems. She is currently guiding students for their research program under V.T.U and other university. Currently, she is working as a Professor and Head of Electronics and Communication Engineering department, Nitte Meenakshi Institute of Technology, Bangalore.

# A Survey of ETL Tools

Mr. Nilesh Mali<sup>1</sup>, Mr. Sachin Bojewar<sup>2</sup>

<sup>1</sup>(Department of Computer Engineering, University of Mumbai, ARMIET, Shahapur, Thane, India)

<sup>2</sup>(Department of Information Technology, University of Mumbai, VIT, Wadala, India)

\*\*\*\*\*

## Abstract:

In most of the organizations valuable data is wasted because of its different formats in various resources. Data warehouses (DWs) are central repositories of integrated data from different disparate sources with an objective to support in decision making process. The Extraction-Transformation-Loading (ETL) processes are the key components of DWs, so the selection of ETL tool is complex and important issue in DWs. ETL is a process of extraction of data, their transformation to desired state by cleaning it, loading it to a target database. This paper first focus the ETL process briefly then discusses analysis of some of the ETL tools based on the generalized criteria for selection of better tool for the improvement of business growth.

**Keywords — Data warehouses, ETL Process, ETL tools, enterprise systems, Business Intelligence.**

\*\*\*\*\*

## I. INTRODUCTION

Data Warehouse (DW) defined by Inmon [1] as “collection of integrated, subject-oriented databases designated to support the decision making process” aims to improve decision process by supplying unique access to several sources. Data warehouses are central repositories of integrated current and historical data from one or more different disparate sources. It mainly contains historical data derived from transaction data and the current data from other sources. It usually characterized by collection of integrated; subject oriented, on-volatile and time variant databases. The heart of DWs is the Extraction-Transformation-Loading (ETL) process. ETL is a process which is used to extract data from various sources, transform that data to desired state by cleaning it and loading it to a target database. The result of this is used to create reports and analyze it. ETL consume up to 70% of resources [3], [5], [4], and [2]. Interestingly [2] reports and

analyses a set of studies proving this fact. ETL is responsible to maintain accuracy and correctness of data.

This paper first focus the ETL process briefly then discusses analysis of some of the ETL tools based on the generalized criteria for selection of better tool. ETL tools are very important for evaluation of Business Intelligence. This includes your results are only as accurate as the input you feed it. Selection of right ETL tool is a fundamental step in achieving your strategic goals.

The research categories of the framework revealed by Nils Schmidt, Mario Rosa, Rick Garcia, Efrain Molina, Ricardo Reyna and John Gonzale[6]. The main purpose to develop a criteria framework to compare the ETL tools against each other.

## II. ETL PROCESS

**ETL (Extract, Transform and Load)** is a process in data warehousing responsible for pulling data

out of different source systems and placing it into a data warehouse.

Basically ETL involves the following tasks:

**A. Extraction**

Extracting the data from different source systems is converted into one consolidated data warehouse format which is ready for transformation purpose.

**B. Transformation**

Transforming the data may involve the following tasks:

- Applying new business rules (so-called derivations, e.g., calculating new dimensions and measures),
- Cleaning the data (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.),
- Filtering the data (e.g., selecting only specific columns to load),
- Splitting a column into multiple columns and merging multiple columns into a column,
- Joining together data from different sources (e.g., lookup, merge),
- Transposing rows into columns and vice versa.
- Applying simple or complex data verification and validation (e.g., if the first 4 columns in a row are empty then reject the row from processing) [7].

**C. Loading**

Loading the data into a data mart or data warehouse or data repository other reporting applications that houses data [7], [8].

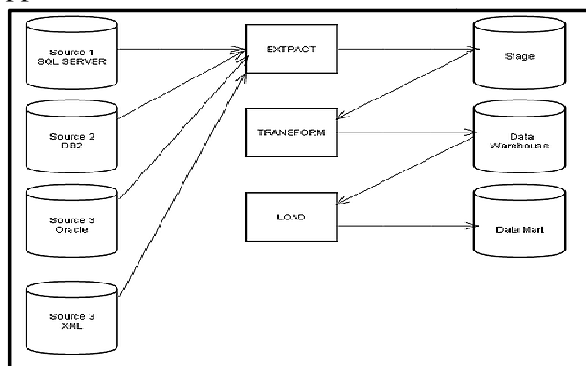


Fig. 1 ETL Workflow

**III. ETL Tools**

The ETL (Extract, transform, load) tools were used to simplify the data management by reducing the absorbed effort. These are designed to save time and money by eliminating the need of 'hand-coding' when a new data warehouse is developed [9]. Depending on the needs of customers there are many types of tools and you have to select appropriate one for you. Most of them are relatively quite expensive, some may be too complex to handle. The most important aspect to start with defining business requirements is selection of right ETL tool. The working of the ETL tools is based on ETL (Extract, transform, load) process.

**A. ETL tools comparison criteria**

- mode of connectivity or adapter support
- mode of data transformation and delivery support
- data modeling and metadata support
- architecture design, development and data governance support
- debugging facility and execution or runtime platform support
- additional services and requirements for vendors
- customers usability support
- cost of hardware or software , installation, OS, Support
- functionality, flexibility and performance support
- infrastructure support

**B. Details of ETL Tools**

Most popular used commercial and freeware (open-sources) ETL Tools are listed below.

**Commercial ETL Tools**

A variety of commercial ETL tools used for data integration, transformation, some of them are listed below

- 1) **IBM** **InfosphereDataStage:** IBMInfoSphereDataStage uses the features of high performance parallel framework and graphical

notation to integrate data across multiple systems. It provides powerful scalable platform for easy and flexible integration of all types of data, including big data at rest (Hadoop-based) or in motion (stream-based), on distributed and mainframe platforms [11]. It manages workload and business rules by optimization of hardware. It is available in various versions such as the Server Edition, the Enterprise Edition, and the MVS Edition. The Enterprise Edition introduces parallel processing architecture and parallel jobs. The Server Edition mainly representing the Server Jobs. The MVS Edition related with mainframe jobs.

**2) InformaticaPowerCenter:**InformaticaPowerCenter is an enterprise data integration platform working as a unit for B2B Data Exchange, Cloud Data Integration, data migration, Complex Event Processing, Data Masking, Data Quality, Data Replication and synchronization, Data Virtualization, Master Data Management, Ultra Messaging, etc.

Basically InformaticaPowerCenter consists of 3 main components.

- InformaticaPowerCenter Client Tools: These are the development tools, installed at developer end to enable the mapping process.
- InformaticaPowerCenter Repository: Repository stores all the metadata for your application so it is the heart of Informatica tools. It act as a data inventory.
- InformaticaPowerCenter Server: Server is responsible for execution of all the data and loading of these data into target system.

**3) Oracle Warehouse Builder (OWB):**Oracle Warehouse Builder (OWB) is an Oracle's ETL tool that enables graphical environment to build, manage and maintain data integration processes in a custom Business Intelligence application. Oracle Warehouse Builder allows creation of dimensional, relational and metadata models, and also star schema data warehouse architectures [9].Oracle

Warehouse Builder supports Oracle Database (releases 8i, 9i and newer) and flat files for target database. It provides data quality, data auditing and full lifecycle management of data and metadata of target database.

**4) Oracle Data Integrator (ODI):**Oracle Data Integrator (ODI) is an ETL based software application used for data transformation and merging or data integration from high-volume, high-performance load, to event-driven, to SOA-enabled data services processes by adding parallelism. The important architecture component of ODI is repository which is collection of all metadata and is accessed by client-server mode or thin client mode. Oracle Data Integrator works as well in the staging and transforming area as the support for other Oracle software[9].

**5) SAS ETL Studio:**This ETL tool is developed by **SAS Enterprise (USA)** which offers an integrated ETL platform. SAS is one of the market leaders which combine data warehousing and intelligence applications for traditional business process. It provides the facility of multithreaded and multiprocessing data extraction to speed up the data transfer and related operations. SAS helps to reduce duplicate or inaccurate data by providing drag and drop interface, not necessary of programming or SQL (Structured Query Language) for managing data [9]. SAS Data Integration Studio enables users to quickly build and edit data integration, to automatically capture and manage standardized metadata from any source, and to easily display, visualize, and understand enterprise metadata and your data integration processes [11],[12].

**6) Business Objects Data Integrator (BODI):** Business Objects Data Integrator is an ETL tool for data integration, mainly focus on data quality features. It also uses data repository for storing of created jobs and projects. BODI is commonly used for building data marts, data warehouses, etc.

It provides a lot of options in data manipulation likes

- Data unification: makes possible quick and trouble-free or easily updating and creating the universes objects.

- Data auditing: data integrity gets verified, especially when the data is read, processed and loaded successfully.
- Data cleansing: maintaining the quality and standard of data.

7) **SQL Server Integration Services (SSIS):** SQL Server Integration Services (SSIS) is a fast and flexible component of the Microsoft SQL Server database software that can be used to perform a wide range of data migration, data transformation and data integration tasks. It can be used to maintain and update multidimensional cube data of SQL Server database for solving complex business problems. SSIS was first released with Microsoft SQL Server in 2005, later replaced Data Transformation Services, which had been a feature of SQL Server from the Version 7.0. It is only available in the editions likes "Standard", "Enterprise" and "Business Intelligence". Using SSIS any type of data can be moves quickly from a variety of source types to a variety of destination types. SSIS includes a rich set of built-in transformations likes Aggregation, Audit, Cache Transform, Data Conversion, Data Mining Query, Dimension Processing, Export & Import Column, Fuzzy Grouping, pivot, row sampling, term extraction, etc. It also provides programmable object model that allows developers to create, store, and load packages for execution. Basically SSIS is used to handle large enterprises, as it requires a Microsoft Server operating system so it requires high operating system and support cost.

1	IBM InfosphereDataStage	<ul style="list-style-type: none"> <li>• flexibility and strongest tool on the market,</li> <li>• provides high level of satisfaction for the clients</li> </ul>	<ul style="list-style-type: none"> <li>• difficult to learn</li> <li>• long and time consuming implementation</li> <li>• requires large amount of memory and processing power</li> </ul>
2	InformaticaPowerCenter	<ul style="list-style-type: none"> <li>• consistent to track the record, easy learning, ability to address real-time data integration schemes</li> <li>• focus on B2B data exchange</li> </ul>	<ul style="list-style-type: none"> <li>• diminishing the value of technologies diminished by several partnerships</li> <li>• In the field experience is limited.</li> </ul>
3	Oracle Warehouse Builder (OWB)	<ul style="list-style-type: none"> <li>• strong, powerful data integration</li> <li>• tight connectivity with respective application</li> <li>• all the tools are integrated in one application of one environment</li> </ul>	<ul style="list-style-type: none"> <li>• focus on ETL solutions only</li> <li>• mostly used for batch-oriented work,</li> <li>• customers are mostly confused</li> </ul>
4	Oracle Data Integrator (ODI)	<ul style="list-style-type: none"> <li>• strong connection to all Oracle data warehousing applications,</li> <li>• all the tools are integrated in one application of one environment</li> </ul>	<ul style="list-style-type: none"> <li>• focus on ETL solutions only</li> <li>• mostly used for batch-oriented work,</li> <li>• Using this future is uncertain</li> </ul>
5	SAS ETL Studio	<ul style="list-style-type: none"> <li>• experienced company, great support and most of all very powerful data integration tool with lots of multi-management features</li> <li>• can work on many operating systems and gather data through number of sources – very flexible</li> </ul>	<ul style="list-style-type: none"> <li>• misplaced sales force, not well recognized organization.</li> <li>• Future is Uncertain.</li> <li>• Cost is high</li> </ul>

TABLE 1. SUMMARY OF THE SURVEY

Sr, No.	Tool	Advantage	Disadvantage
---------	------	-----------	--------------



		<ul style="list-style-type: none"> <li>• great support for the business-class companies as well for those medium and minor ones</li> </ul>	
6	Business Objects Data Integrator (BODI)	<ul style="list-style-type: none"> <li>• SAP integration</li> <li>• Better data modeling and data-management support;</li> <li>• provides SAP tool for data mining</li> <li>• Quick learning and ease to use</li> </ul>	<ul style="list-style-type: none"> <li>• different companies uses different SAP Business Objects.</li> <li>• uncertain future.</li> <li>• not supported as a stand-alone capable application of few organization.</li> </ul>
7	Microsoft SQL Server Integration Services(SSIS)	<ul style="list-style-type: none"> <li>• integrates data with standard</li> <li>• Ease speed of implementation details.</li> <li>• low cost, excellent support and distribution</li> </ul>	<ul style="list-style-type: none"> <li>• Window limitation, complexity increases</li> <li>• unclear strategy and vision</li> </ul>

### Freeware (open-sources) ETL Tools

A variety of open source ETL tools used for data integration, transformation, some of these listed below:

1) **Pentaho Data Integration (Kettle)**: Kettle has been recently acquired by the Pentaho group so it is also called as Pentaho Data Integration. It is a leading open source ETL application or tool in the market. Basically it is composed of four elements, **ETTL** (**extraction** from different sources, **transport** of data, **transformation** of data, **loading** of data into data warehouse). Kettle can supports deployment on single node computer as well as on a cloud, or cluster. It can load and process big data sources in familiar ways [10] by providing flexibility and security. Kettle can support various data sources

and databases like oracle, MySQL, AS/400, MS Access, MS SQL Server, Sybase, IBM DB2, dBase, Informix, Hypersonic, any database using ODBC on Windows, etc [8]. By purchasing the Enterprise Edition would provide the services like recovery and backup. Pentaho is easy to use as it being open source software allows administrators to utilize its new drag and drop environment.

The main components of Pentaho Data Integration are:

**Spoon** : Related with the transformation (reading, validating, refining and writing) of data.

**Pan** : Used to run data transformations designed in Spoon.

**Chef** : Used to automate the database update process

**Kitchen** : Used to help and to execute the jobs in a batch mode.

**Carte** : a web server which allows remote monitoring of the running Pentaho Data integration ETL processes through a web browser [8].

2) **Talend Open Studio**: Talend Open Studio is a Data Integration platform that enables designing of data integration processes and their monitoring and operates as a code generator, producing data-transformation scripts and underlying programs in Java. It consists of metadata repository which provides the data (definitions and configuration related data for each job) to all the components of Talend Open Studio. It is commonly used for data migration, synchronization or replication of databases. It also used to improve the quality of big data. It easy to operate does not require extra technical skills.

3) **Clover ETL**: Clover is a Commercial Open Source ETL tool considered for data transformation and integration, cleanse, and distribute data into applications, databases, and data warehouses. The Clover ETL tool is based on Java so it is independent and resource- efficient. It can be used as standalone as well as a command- line application or server application or just like Java library it can be even embedded in other

applications [9]. Clover ETL is operated on any type of operating system like windows, Linux, HP-UX, etc. It can be both used on low-cost PC as on high-end multi processors servers [9]. Graph or transformation dataflow are used to represent data flow in Clover ETL. Edges of graph can represent the data flow from one component to another.

#### Clover ETL components

- **CloverETLEngine** : the core for running data transformation graphs
- **CloverETLDesigner**: a commercial visual data integration tool used to design and execute transformation graphs
- **Clover ETL Server** :An enterprise automation and data integration monitoring platform having the features such as workflows, scheduling, monitoring, user management, or real-time ETL abilities.

**4) Jasper ETL:** Jasper ETL is one of the easiest open sources ETL tool considered for data integration, cleansing, transformation and movement on the market [13]. In this case it is said to reduce both the costs of ownership and the complexity of an IT infrastructure services. It consists of aggregation or integration of large data from different data sources. While transforming the data it maintains both consistency and accuracy of data. Finally loads the data into data warehouse in optimize way. This tool is easily affordable to anyone. It is easy to manage and proven superior performance than many commercial ETL tools. Is can be used to any type of business, it may be small or complex.

#### IV. CONCLUSIONS

ETL tools are designed and used to save time and cost when a new data mart or data warehouse is developed. ETL tools are basically at the foundation of Business Intelligence for migration or transformation of Data from one format to another or Data mobility then you might have to employ

these ETL tools to enable your business process. They help to extract the data from different heterogeneous database, to transform the data into a unified standard format by cleansing and applying various processes and finally load it into data mart or data warehouse. From our survey we have studied different commercial ETL software tools; we find that Microsoft SQL Server Integration Services (SSIS) are mostly satisfied the needs of large organizations, as it can handle the large database. In case of freeware or open sources ETL tools, Pentaho Data Integration (Kettle) is mostly used for small enterprises, as it limits the speed and having limited debugging facility. This survey mostly helps us for selection of best ETL tool, but ultimately, the decision will depends on your organization and factors considered for selection of best ETL tool.

#### ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who contributed in some way to the work described in this paper. My sincere thanks to my project guide **Prof. SachinBojewar** for giving me intellectual freedom of work and guiding me on proper time. I would also like to thanks head of computer department and to the principal of AlamuriRatnamala Institute of engineering & technology, Shahapur, Thane for extending his support for our work.

#### REFERENCES

- [1] W. Inmon D. Strauss and G.Neushloss, "DW 2.0 The Architecture for the next generation of data warehousing", Morgan Kaufman, 2007.
- [2] A.Simitisis, P. Vassiliadis, S.Skiadopoulos and T.Sellis "Data Warehouse Refreshment", Data Warehouses and OLAP: Concepts, Architectures and Solutions, IRM Press, 2007, pp 111-134.
- [3] R. Kimball and J. Caserta. "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", Wiley Publishing, Inc, 2004.
- [4] A. Kabiri, F. Wadjinny and D. Chiadmi, "Towards a Framework for Conceptual Modelling of ETL Processes", Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science Volume 241, pp 146-160.
- [5] P. Vassiliadis and A. Simitisis, "EXTRACTION, TRANSFORMATION,

ANDLOADING“, [http://www.cs.uoi.gr/~pvassil/publications/2009\\_DB\\_encyclopedia/Extract-Transform-Load.pdf](http://www.cs.uoi.gr/~pvassil/publications/2009_DB_encyclopedia/Extract-Transform-Load.pdf)

Computer Interaction,  
Programming Languages,  
Management Information System,  
Software Architecture, etc.

- [6] Nils Schmidt, Mario Rosa, Rick Garcia, Efrain Molina, Ricardo Reyna and John Gonzale, “ETL TOOL EVALUATION- A Criteria Framework “
- [7] Data Warehouse, <http://datawarehouse4u.info> accessed on September 10, 2014
- [8] ETL Tools information, [http://etl-tools.info/en/bi/etl\\_process.htm](http://etl-tools.info/en/bi/etl_process.htm) accessed on September 12, 2014
- [9] ETL Tools, <http://www.etltools.net> accessed on September 12, 2014
- [10] Pentaho data integration, <http://www.pentaho.com/product/data-integration> accessed on September 14, 2014
- [11] IBM InfosphereDataStage, [www.ibm.com/software/products/en/ibminfodata](http://www.ibm.com/software/products/en/ibminfodata) accessed on September 14, 2014
- [12] SAS ETL studio, <http://support.sas.com/software/products/etls/> accessed on September 14, 2014
- [13] Jasper ETL tool, <http://community.jaspersoft.com> accessed on September 14, 2014

## **AUTHORS**

Mr. NileshMali is currently a graduate student pursuing masters in Computer Engineering at ARMIE, Thane, University of Mumbai, India. He has received hisB.E in Computer Engineering from University of Pune. He has 3 year of past experience in teaching. His areas of interest are distributed database, data mining, Information Retrieval, etc.



Mr. SachinBojewar is Associate professor in Information Technology Department, at VIT, Wadala, University of Mumbai, India. He has completed his B.E and M.E in computer Engineering. Currently He is pursuing his PhD in Software Engineering. He is having 25 years of experience in teaching. His areas of interest are Object Oriented Design, Software Engineering, Management Information System, Human



# Towards measuring learner's concentration in E-learning systems

Saoutarrih Marouane<sup>1</sup>, Sedrat Najlaa<sup>2</sup>, Tahiri Abderrahim<sup>3</sup>, Elkadiri Kamal Eddine<sup>4</sup>  
<sup>1,2,3,4</sup>(LIROSA Laboratory Abdelmalek Essaâdi University, Morocco)

\*\*\*\*\*

## Abstract:

Due to the advantages of e-learning platforms, many universities and schools around the world had adopted e-learning platforms in their educational systems. Students appreciate using e-learning platforms because of the advantages that offers like time flexibility, location flexibility and fast access to courses documents. However the major problem that they have with e-learning platforms is the difficulty to maintain concentration and focus during the learning time. The absence of teacher leads students to feel free to use phones, navigate in web sites, and play video games and by consequence they lose their focus and they couldn't understand the course.

In the traditional way of learning, teachers are face to face with students so they know how to keep them engaged in a lesson and how to get them back on track when they lose focus. Those ways used by teachers to keep students engaged are difficult to be implemented in e-learning systems. That's why many studies had aimed this problem and try to find the best ways to implement those methods to maintain learners' concentration during the use of e-learning platforms. There are many solutions under studies such as using cameras, brain waves, blood oxygen, heartbeat and blood pressure to detect when students are losing focus to activate an action to get the focus back. In this paper we will focus on the different methods for measuring students' level of concentration

**Keywords — E-learning, concentration, skin temperature, web camera, EEG signals.**

\*\*\*\*\*

## I. INTRODUCTION

Compared to traditional face to face learning [1], E-learning had offered either for teachers and students many advantages. E-learning gives teachers possibility to share courses, quizzes and multimedia files. They can also share online discussion with students which can lead to build good relationship between teachers and students. Furthermore, E-learning offers to students more flexibility to consult courses anywhere with a considerable flexibility of time, it also reduces travel time and travel costs for off-campus students. In addition, e-learning allows students to retain more information than with traditionally instructor led training due to the variety of elements combined to e-learning such as videos, audios, simulations or webinars that reinforce the learning message, [2] Despite all the advantages of E-learning, e-learning

platforms had neglected a very important element in the learning operation which is the control of students' intention during the use of the platform in order to keep students engaged.

We need to keep in mind that when students are using e-learning platforms, there is no teacher to supervise and alarm the student to give more intention to the course, so the learner can quickly feel bored and sleepy and by consequence learner find a difficulty to take the course.

Many researches are interested by this problem and try to find solution to control students' intention by analysing their behaviour and motivation using cameras, brain waves, blood oxygen, heartbeat and blood pressure.

In this paper we will present different measuring methods of learner concentration. The organization of this paper will be as follows. Section 2

introduces e-learning evolution over the time. Section 3 shows the relation between a successful e-learning and learner concentration. Section 4 presents different methods of measuring learner concentration level. The last section concludes the paper and provides future directions for further research.

## **II. E-LEARNING EVOLUTION**

E-learning can be defined by the use of telecommunication and technology to deliver education and training knowledge. Before 1983, As stated by Kiffmeyer(2004), the learning method was Instructor-Led Training (ILT) in which learners can interact with teachers and with themselves, this way of learning needed more costs for travelling and lodging. After 1984, with the apparition of computers, PowerPoint and cd-rooms, more transportability had been given to courses which makes students free to consult their courses anytime and anywhere, and by consequence they can save learning time and cost. The disadvantage of this way of learning was the lack of student-instructor and student-student interaction. This problem had been solved after thanks to the apparition of web technology in 1994. With streaming, media players, mailing, etc. The course had become more illustrated and more interactive. The problem of this method was the web additional costs due to the low band-width and the high price of internet access.

After 2000, with the evolution of web technologies, network, band-width access and lowering of internet cost, learning had become in live with a good interaction teacher-student and student-student.

## **III. SUCCESSFUL E-LEARNING AND LEARNERS CONCENTRATION**

To benefits from advantages of e-learning evolution and to have a successful e-learning, there must be a good learner satisfaction. Because when a learner is satisfied his intention increases to continue using the e-learning course. To ameliorate learner satisfaction many studies had been established. Sun, Tsai, Finger Chen and Yeh(2008)

had done an empirical investigation of the critical factors influencing learner satisfaction, and had found that factors influencing learner satisfaction can be categorized into six dimensions:

learner dimension, instructor dimension, course dimension, technology dimension, design dimension and environment dimension. Controlling these factors can considerably increase students' satisfaction. But even if all those factors are taken into consideration in the e-learning system, student can lose his concentration during the course because of familiar or health problems, the use of mobile phone during the course or browsing web pages for fun during learning. Knowing that teachers cannot grasp the real learning situation during e-learning courses, students will not follow course evolution and as a result they will learn nothing. Therefore, it's very important to find a method that allows instructor to know student concentration level and by consequence alarming him about his concentration level.

## **IV. METHODS OF MEASURING LEARNER CONCENTRATION**

Many researches are interested in controlling learner concentration during learning and had developed several methods to measure it.

### **A. Skin temperature indicator**

According to an experiment done by Nomura, Hasegawa-Ohira, Korosawa, Hanasaka, Yajima and Fukumura (2012), Skin temperature (ST) is one of the indicators that can be used to detect student involvement in e-learning courses.

ST variation reflects changes in blood stream in a non-stressing behavior. Having a close relationship between blood stream and "the choice of behaviors", a variation of skin temperature implies a change in student use of e-learning materials.

Nomura, Hasegawa-Ohira, Korosawa, Hanasaka, Yajima and Fukumura (2012) had used integrated thermistor into the mouse to measure student temperature without stressing users, the analysis of the values measured combined with real observation of student intention had shown that



there is a relationship between skin temperature variation and students intention.

### **B. Visual attention recognition**

Visual attention recognition (VAR) is one of the most used ways to measure user's intention. The first step of VAR is to capture each second user's pictures. The second step is to analyze them with software to detect students that had loss concentration.

According to a VAR experiment done by Rakahashi and Arita (2014), the accuracy of results is about 83%.

### **C. EEG Signal for attention recognition**

EEG signals can also reflect user intention by measuring human brain signals with mobile brainwave sensors. The use of this indicator to determine level concentration of students is viable because humans in general cannot control their fluctuations in their EEG signal (Ning-Han, Cheng-Yu and Hsuan-Chen, 2013).

According to an experiment done by Ning-Han, Cheng-Yu and Hsuan-Chin (2013), the accuracy of this method is about 76.82%.

•

## **V. CONCLUSIONS**

Measuring student's concentration during learning is very important to ameliorate learning quality and efficiency; many research works had found methods to measure student's intention. Those methods can be explored in further researches to find suitable methods to alarm students about their concentration level without stressing them.

## **REFERENCES**

- [1] Benta,D. Bologa,G. Dzitac,I.(2014). "E-learning Platforms in Higher Education", Case Study.Procedia Computer Science Volume 31, 2014, Pages 1170–1176.
- [2] Institute for Interactive Technologies (2006). "E-Learning Concepts and Techniques", Bloomsburg University of Pennsylvania, USA.
- [3] Kiffmeyer, Michael. (2004, November 9). "The evolution of e-learning", Retrieved April 21, 2006 from <http://knowledgemanagement.ittoolbox.com/documents/popular-q-and-a/the-evolution-of-elearning-2902>.
- [4] Ning-Han ,L. Cheng-Yu, C. Hsuan-Chin, C.(2013). "Recognizing the Degree of Human Attention Using EEG Signals from Mobile Sensors", Sensors 2013, 13, 10273-10286.
- [5] Nomura,S. Hasegawa-Ohira,M. Kurosawa,Y. Hanasaka,Y. Yajima,K. Fukumura,Y. (2012). "SKIN TEMPERETURE AS A POSSIBLE INDICATOR OF STUDENT'S INVOLVEMENT IN E-LEARNING

SESSIONS", Nagaoka University of Technology, Japan. International Journal of Electronic Commerce Studies Vol.3, No.1, pp.101-110, 2012.

- [6] Rakahashi,K.Arita, K.(2014)."Improvement of detection for warning students in e-learning using web cameras", Hiroshima City University, Japan. Procedia Computer Science Volume 35, 2014, Pages 747–756.
- [7] Richard, K-L. (2004)."E-learning compared with face to face: Differences in the academic achievement of postgraduate business student", Australasian Journal of Educational Technology 2004, 20(3), 316-336.
- [8] Sun, P.C., Tsai, R.J., Finger, G., Chen, Y.Y. & Yeh, D. (2008)." What Drives a Successful E-Learning? An Empirical Investigation of the Critical Factors Influencing Learner Satisfaction", Computers & Education, 50(4), 1183-1202.

# A Study on Behavioural Malware Detection by Using Delay Tolerant Networks

K.Ravikumar<sup>1</sup>, V. Vinothkumar<sup>2</sup>

<sup>1</sup>( Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu),Thanjavur)

<sup>2</sup> (Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.)

\*\*\*\*\*

## Abstract:

The delay-tolerant-network (DTN) model is becoming a via communication alternative to the traditional infrastructural model for modern mobile consumer electronics equipped with short-range communication technologies such as Bluetooth, NFC, and Wi-Fi Direct. Proximity malware is a class of malware that exploits the opportunistic contacts and distributed nature of DTNs for propagation. Behavioral characterization of malware is an effective alternative to pattern attaching in detecting malware, especially when dealing with polymorphic or obfuscated malware. In this paper, we first propose a general behavioral characterization of proximity malware which based on Naive Bayesian model. We identify two unique challenges for extending Bayesian malware detection to DTNs and propose a simple yet effective method, look-ahead, to address the challenges. Furthermore, we propose two extensions to look-ahead, dogmatic filtering and adaptive look-ahead, to address the challenge of “malicious nodes sharing false evidence”. Real mobile network traces are used to verify the effectiveness of the proposed methods.

**Keywords — DTN,WiFi.**

\*\*\*\*\*

## I. INTRODUCTION

A delay-tolerant network is a network designed to operate effectively over extreme distances such as those encountered in space communications or on an interplanetary scale. In such an environment, long latency -- sometimes measured in hours or days -- is inevitable.

The popularity of mobile consumer electronics, like laptop computers, PDAs, and more recently and prominently, smart phones, revives the delay-tolerant-network (DTN) model as an alternative to the traditional infrastructure model.

The widespread adoption of these devices, coupled with strong economic incentives, induces a class of malware that specifically targets DTNs. We call this class of malware proximity malware. Proximity malware based on the DTN model brings unique security challenges that are not present in the infrastructure model. In the infrastructure model, the cellular carrier centrally monitors

networks for abnormalities moreover the resource scarcity of individual nodes limits the rate of malware propagation.

A prerequisite to defending against proximity malware is to detect it. In this paper, we consider a general behavioural characterization of proximity malware. Behavioural characterization, in terms of system call and program flow, has been previously proposed as an effective alternative to pattern matching for malware detection. In our model, malware-infected nodes behaviours are observed by others during their multiple opportunistic encounters: Individual observations may be imperfect, but abnormal behaviours of infected nodes are identifiable in the long-run.

### 1.1.OBJECTIVE

Network is the combination of Nodes. Each node will communicate with its neighbors and share their data. If a node is affected by a malware it's necessary to clear it else its neighbors will communicate with it and they also affected by

malware. Hence detection of malware is important. Here we discuss some methods for the detection of malware.

## **2. EXISTING SYSTEM**

Previous researches quantify the threat of proximity malware attack and demonstrate the possibility of launching such an attack, which is confirmed by recent reports on hijacking hotel Wi-Fi hotspots for drive-by malware attack.

With the adoption of new short-range communication technologies such as NFC and Wi-Fi Direct that facilitate spontaneous bulk data transfer between spatially proximate mobile devices, the threat of proximity malware is becoming more realistic and relevant than ever. Proximity malware based on the DTN model brings unique security challenges that are not present in the model.

### **2.1. EXISTING SYSTEM DISADVANTAGES**

- Central monitoring and resource limits are absent in the DTN model.
- Very risk to collecting evidence and also having insufficient evidence.
- It is filter the false evidence in sequentially and distributed.

## **3. PROPOSED SYSTEM**

Behavioral characterization, in terms of system call and program flow, has been previously proposed as an effective alternative to pattern matching for malware detection. In our model, malware-infected nodes' behaviors are observed by others during their multiple opportunistic encounters: Individual observations may be imperfect, but abnormal behaviors of infected nodes are identifiable in the long-run. We identify challenges for extending Bayesian malware detection to DTNs, and propose a simple yet effective method, look-ahead, to address the challenges. Furthermore, we propose two extensions to look-ahead, dogmatic filtering and adaptive look-ahead, to address the challenge of "malicious nodes sharing false evidence".

### **3.1. PROPOSED SYSTEM ADVANTAGES**

- Real mobile network traces are used to verify the effectiveness of the proposed methods.
- The proposed evidence consolidation strategies in minimizing the negative impact of liars on the shared evidence's quality.
- It is used to identify the abnormal behaviors of infected nodes in the long-run.

### **3.2. PROPOSED TECHNIQUES**

Proximity malware is a malicious program that disrupts the host node's normal function and has a chance of duplicating itself to other nodes during (opportunistic) contact opportunities between nodes in the DTN. When duplication occurs, the other node is infected with the malware. We present a general behavioral characterization of proximity malware, which captures the functional but imperfect nature in detecting proximity malware.

Under the behavioral malware characterization, and with a simple cut-off malware containment strategy, we formulate the malware detection process as a distributed decision problem. We analyze the risk associated with the decision, and design a simple, yet effective, strategy, look-ahead, which naturally reflects individual nodes' intrinsic risk inclinations against malware infection.

We present two alternative techniques, dogmatic filtering and adaptive look-ahead, that naturally extend look-ahead to consolidate evidence provided by others, while containing the negative effect of false evidence. A nice property of the proposed evidence consolidation methods is that the results will not worsen even if liars are the majority in the neighborhood

### **3. METHODOLOGIES**

Methodologies are the process of analyzing the principles or procedure for behavioral characterizing of node with two methods, dogmatic filtering and adaptive look-ahead, for consolidating evidence provided by other nodes, while containing the negative impact of liars in delay tolerant network.

### 3.1.ADVANTAGES

- Real mobile network traces are used to verify the effectiveness of the proposed methods.
- The proposed evidence consolidation strategies in minimizing the negative impact of liars on the shared evidence's quality.
- It is used to identify the abnormal behaviors of infected nodes in the long-run.

### MODULES

- I.Authentication
- II.Network Nodes
- III.Malware Detection
- IV.Evidence Analysis
- V.Evil Node Revocation

### 3.2.MODULE DESCRIPTION

#### Authentication

If you are the new user going to consume the service then they have to register first by providing necessary details. After successful completion of sign up process, the user has to login into the application by providing username and exact password. The user has to provide exact username and password which was provided at the time of registration, if login success means it will take up to main page else it will remain in the login page itself..

#### Network Nodes

Under this module, the network nodes which are interconnected by local area network, that node ip address will be fetched in order to share the resources among the network. As well as the performance of individual system have been analyzed to assess the behavior

#### Malware Detection

Malware detection module helps to identify the evil node which is affected by malware program

#### Evidence Analysis

This module used to investigate about evidences of nodes by collecting assessments before a normal node get affected by malware program. Evidence aging process helps to discard outdated assessments of a node and evidence consolidation helps to filter negative assessments of a node provided by the other nodes.

#### Evil Node Revocation

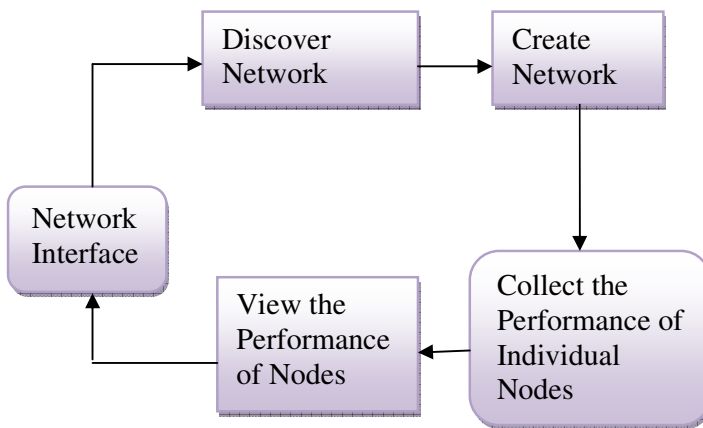
After detection of evil node, we need to drop the communication with that in order to prevent from malware spreading and the evil node details are transferred to database for further reference. Finally evil node gets revoked from the network computer list.

### 4.CONCLUSION

Our system proposes a general behavioral characterization of DTN-based proximity malware. We present dogmatic filtering and adaptive look-ahead technique to address two unique problems “insufficient evidence vs. evidence collection risk” and “filtering false evidence sequentially and distributed”.

### 5.REFERENCES

- 1.Y. Li, P. Hui, L. Su, D. Jin, and L. Zeng, “An optimal distributed malware defense system for mobile networks with heterogeneous devices,” in Proc. IEEE SECON, 2011.
2. D. Dash, B. Kveton, J. Agosta, E. Schooler, J. Chandrashekar, A. Bachrach, and A. Newman, “When gossip is good: Distributed probabilistic



inference for detection of slow network intrusions,” in Proc. AI, 2006.

3. F. Li, Y. Yang, and J. Wu, “CPMC: an efficient proximity malware coping scheme in Smartphone-based mobile networks,” in Proc. IEEE INFOCOM, 2010.

4. U. Bayer, P. Comparetti, C. Hlauschek, C. Kruegel, and E. Kirda, “Scalable, behavior-based malware clustering,” in Proc. IEEE NDSS, 2009.

5. J. Su, K. Chan, A. Miklas, K. Po, A. Akhavan, S. Saroiu, E. de Lara, and A. Goel, “A preliminary investigation of worm infections in a bluetooth environment,” in Proc. ACM WORM, 2006.

6. S. Cheng, W. Ao, P. Chen, and K. Chen, “On modeling malware propagation in generalized social networks,” IEEE Comm. Lett., vol. 15, no. 1, pp. 25–27, 2011.

7. S. Kamvar, M. Schlosser, and H. Garcia-Molina, “The eigentrust algorithm for reputation management in P2P networks.” in Proc. ACM WWW, 2003.

8. A. Bose and K. Shin, “On mobile viruses exploiting messaging and bluetooth services,” in Proc. IEEE SecureComm, 2006.



# Sybil Belief: A Semi- Creation New Approach for Structure –Based Sybil Detection

K.Ravikumar<sup>1</sup>, B. Selvam<sup>2</sup>

<sup>1</sup>(Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu), Thanjavur)

<sup>2</sup> (Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.)

\*\*\*\*\*

## Abstract:

Sybil attacks are a fundamental threat to the security of distributed system. There has been a growing interest in leveraging social network to mitigate Sybil attacks. We introduce Sybil belief a semi supervised learning framework to detect Sybil nodes. Sybil Belief takes a social network of the nodes in the system, a small set of known benign nodes, and, optionally, a small set of known Sybil's as input. We show that Sybil Belief is able to accurately identify Sybil nodes with low false positive rates and low false negative rates. Sybil Belief is resilient to noise in our prior knowledge about known benign and Sybil nodes. Sybil accounts in online social networks are used for criminal activities such as spreading spam or malware stealing other users' private information and manipulating web search results. Sybil defenses require users to present trusted identities issued by certification authorities. However, such approaches violate the open nature that underlies the success of these distributed systems.

**Keywords — Sybil attack, social work, Detection.**

\*\*\*\*\*

## I. INTRODUCTION

Sybil attacks, where a single entity emulates the behavior of multiple users, form a fundamental threat to the security of distributed systems .Example systems include peer-to peer networks, email, reputation systems, and online social networks. For instance, in 2012 it was reported that 83 million out of 900 million Face book accounts are Sybils. Sybil accounts in online social networks are used for criminal activities such as spreading spam or malware, stealing other users' private information , and manipulating web search results via "+1" or "like" clicks. Traditionally, Sybil defenses require users to present trusted identities issued by certification authorities. However, such approaches violate the open nature that underlies the success of these distributed systems. Recently, there has been a growing interest in leveraging social networks to mitigate Sybil attacks. These schemes are based on the observation that, although an attacker can create arbitrary Sybil users and social connections among themselves, he or she can only establish a limited number of social

connections to benign users. As a result, Sybil users tend to form a community structure among themselves, which enables a large number of Sybil users to integrate into the system. Note that it is crucial to obtain social connections that represent trust relationships between users, otherwise the structure-based Sybil detection mechanisms have limited detection accuracy.

### 1.1. OBJECTIVE

Sybil Belief a semi-supervised learning framework, to perform both Sybil classification and Sybil ranking. Sybil Belief overcomes a number of drawbacks of previous work. We extensively evaluate the impact of various factors including parameter settings in Sybil Belief, the number of labels, and label noise on the performance of Sybil Belief using synthetic social networks.

### 1.2. EXISTING SYSTEM

Sybil detection mechanisms rely on the assumption that the benign region is fast mixing. we recast the problem of finding Sybil users as a semi-supervised learning problem, Sybil detection

methods decrease dramatically when the benign region consists of more and more communities they cannot tolerate noise in their prior knowledge about known benign or Sybil nodes and they are not scalable.

## **2.1.EXISTING SYSTEM DISADVANTAGES**

- They can bootstrap from either only known benign or known Sybil nodes limiting their detection accuracy.
- They are not scalable

## **PROPOSED SYSTEM**

We analyze the problem of behavioral characterization Sybil accounts in online social networks are used for criminal activities such as spreading spam or malware.

## **2.2. PROBLEM DEFINITION**

Sybil Belief using both synthetic and real world social network topologies. We show that Sybil Belief is able to accurately identify Sybil nodes with low false positive rates and low false negative rates. Sybil Belief is resilient to noise in our prior knowledge about known benign and Sybil nodes. They can bootstrap from either only known benign or known Sybil nodes, limiting their detection accuracy, they cannot tolerate noise in their prior knowledge about known benign or Sybil nodes. They are not scalable.

## **2.3. METHODOLOGIES**

Sybil Belief a semi-supervised learning framework, to perform both Sybil classification and Sybil ranking. Sybil Belief overcomes a number of drawbacks of previous work. We extensively evaluate the impact of various factors including parameter settings in Sybil Belief, the number of labels, and label noise on the performance of Sybil Belief using synthetic social networks.

### **2.3.1. MODULES**

#### **USER**

- 1.Authentication
- 2.Registration Form
- 3.Edit profile
- 4.Request Friend

### **2.3.2 MODULE DESCRIPTION**

#### **Authentication**

If you are the new user going to consume the service then they have to register first by providing necessary details. After successful completion of sign up process, the user has to login into the application by providing username and exact password. The user has to provide exact username and password which was provided at the time of registration, if login success means it will take up to main page else it will remain in the login page itself..

#### **Registration Form:**

In this Module If he is a new user he needs to enter the required data to register the form and the data will be stored in server for future authentication purpose.

#### **Edit profile:**

A profile can be used to store the description of the characteristics of person. This information can be exploited by systems taking into account the persons' characteristics and preferences.

#### **Request Friend:**

Someone is the act of sending another user a friend request on Social Network. The two people are social network friends once the receiving party accepts the friend request. Deleting a friend request removes the request.

#### **Blocked Profile:**

Admin can compare with profile to all profile. It will match the profile. This profile is identifying fake profile. So that profile will be blocked.

## **4.CONCLUSION**

In this paper, we propose Sybil Belief, a semi-supervised learning framework, to detect Sybil nodes in distributed systems. Sybil Belief takes social networks among the nodes in the system, a small set of known benign nodes, and, optionally, a small set of known Sybil nodes as input, and then Sybil Belief propagates the label information from the known benign and/or Sybil nodes to the remaining ones in the system. We

extensively evaluate the influence of various factors including parameter settings in the SybilBelief, the number of labels, and label noises on the performance of SybilBelief. Moreover, we compare SybilBelief with state-of-the-art Sybil classification and ranking approaches on real-world social network topologies. Our results demonstrate that SybilBelief performs orders of magnitude better than previous Sybil classification mechanisms and significantly better than previous Sybil ranking mechanisms. Furthermore, SybilBelief is more resilient to noise in our prior knowledge about known benign nodes and known Sybils. Interesting avenues for future work include evaluating Sybil-Belief and previous approaches with datasets containing real Sybils and applying our SybilBelief framework to other security and privacy problems such as graph based Botnet detection, reputation systems, and private information inference.

## 5. REFERENCES

- [1] J. R. Douceur, "The Sybil attack," in IPTPS, 2002.
- [2] Malicious/fake accounts in Facebook, <http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/index.html>.
- [3] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in IEEE S & P, 2011.
- [4] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networks," in WWW, 2009.
- [5] P. L. Fong, "Preventing Sybil attacks by privilege attenuation: A design principle for social network systems," in IEEE S & P, 2011. [6] Google Explores +1 Button To Influence Search Results, "http://www.tekgoblin.com/2011/08/29/google-explores-1-button-toinfluence-search-results/."
- [6] Google Explores +1 Button To Influence Search Results,
- "http://www.tekgoblin.com/2011/08/29/google-explores-1-button-toinfluence-search-results/."
- [7] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove, "An analysis of social network-based Sybil defenses," in SIGCOMM, 2010.
- [8] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman., "SybilGuard: Defending against Sybil attacks via social networks," in SIGCOMM, 2006.
- [9] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "SybilLimit: A nearoptimal social network defense against Sybil attacks," in IEEE S & P, 2008.
- [9] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "SybilLimit: A nearoptimal social network defense against Sybil attacks," in IEEE S & P, 2008.
- [10] G. Danezis and P. Mittal, "SybilInfer: Detecting Sybil nodes using social networks," in NDSS, 2009.

# Multi Agent System in Distributed Agent Network

K.Ravikumar<sup>1</sup>, A. Surendar<sup>2</sup>

(<sup>1</sup>Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu), Thanjavur.)

(<sup>2</sup>Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.)

\*\*\*\*\*

## Abstract:

Cloud computing is a working; process involves virtualization, distributed computing, networking, software and web services. A cloud consists of several elements such as clients, datacenter and distributed servers. It includes fault tolerance, high availability, scalability, flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc. Cloud computing deliver the computing as a services whereby share resources, software, information via Internet which are accessed by the browser. The business software and data are stored in server at Remote Location, Cloud computing provides the kinds of services that are Infrastructure, Software, platform as a services. Gossip Protocol is effective protocol for the dynamic load balance in the distributed system and continuously execute process input & output process. Resources allocation policies are computed by protocols. Our contribution includes outlining distributed middleware architecture and presenting one of its key elements: a gossip protocol that (1) ensures fair resource allocation among sites, (2) dynamically adapts the allocation to load changes and (3) scales both in the number of physical machines and applications. The protocol continuously executes on dynamic, local input and does not require global synchronization, as other proposed gossip protocols.

**Keywords — Agent, User Interface.**

\*\*\*\*\*

## 1. INTRODUCTION:

Cloud Computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing data storage, processing and bandwidth. The cloud computing model is comprised of a front end and a back end. These two elements are connected through a network, in most cases the Internet. The front end is the vehicle by which the user interacts with the system; the back end is the cloud itself. The front end is composed of a client computer, or the computer network of an enterprise, and the applications used to access the cloud. The back end provides the applications, computers, servers, and data storage that creates the cloud of services.

We propose a gossip protocol that ensures fair resource allocation among applications, dynamically adapts the allocation to load changes and scales both in the number of physical machines and sites.

## 2. Existing System:

Application placement in datacenters is often modeled through mapping a set of applications onto a set of machines such that some utility function is maximized under resource constraints. This approach has been taken, and solutions from these works have been incorporated in middleware products. The problem of resource management is application placement and load balancing in processor networks.

### 2.1.EXISTING SYSTEM

Sybil detection mechanisms rely on the assumption that the benign region is fast mixing. we

recast the problem of finding Sybil users as a semi-supervised learning problem, Sybil detection methods decrease dramatically when the benign region consists of more and more communities they cannot tolerate noise in their prior knowledge about known benign or Sybil nodes and they are not scalable.

Dynamic resource management for a large-scale cloud environment is problematic one. We propose a gossip protocol that ensures fair resource allocation among sites/applications, dynamically adapts the allocation to load changes and scales both in the number of physical machines and sites/applications. We present a protocol that computes an optimal solution without considering memory constraints and prove correctness and convergence properties. Next, we extend that protocol to provide an efficient heuristic solution for the complete problem, which includes minimizing the cost for adapting an allocation. The protocol continuously executes on dynamic, local input and does not require global synchronization, as other proposed gossip protocols do

### **3. PROBLEM DEFINITION**

Our contribution includes outlining distributed middleware architecture and presenting one of its key elements: a gossip protocol that ensures fair resource allocation among sites/applications, dynamically adapts the allocation to load changes and scales both in the number of physical machines and sites/applications. The protocol continuously executes on dynamic, local input and does not require global synchronization, as other proposed gossip protocols

#### **3.1. METHODOLOGY:**

##### **3.1.1. Gossip Protocol perform:**

Gossip-based protocols have recently gained notable popularity. Apart from traditional applications for database replication gossiping algorithms have been applied to solve numerous other practical problems including failure detection, resource monitoring and data aggregation.

##### **3.1.2. Advantage**

- 1. Continuous execution is possible and process
- 2. Suitable for heavy execution processes.
- 3. Resource allocation handling is possible as this work.
- 4. No need for the global synchronization because the protocol gossip already capability for load balancing

#### **4. USER INTERFACE DESIGN:**

In this module we design the windows for the project. These windows are used to send a message from one peer to another. We use the Swing package available in Java to design the User Interface. Swing is a widget toolkit for Java. It is part of Sun Microsystems' Java Foundation Classes (JFC) — an API for providing a graphical user interface (GUI) for Java programs. In this module mainly we are focusing the login design page with the Partial knowledge information. Application Users need to view the application they need to login through the User Interface GUI is the media to connect User and Media Database and login screen where user can input his/her user name, password and password will check in database, if that will be a valid username and password then he/she can access the database.

#### **4.1. CLIENT APPLICATION COMMUNICATION -PROTOCOL**

Distributed middleware architecture. Distributed protocols are gossip protocols, specifically. We presented architecture and a generic gossip protocol for application placement in a cloud environment. The protocol can be instantiated for specific management objectives. It computes a distributed heuristic solution to an optimization problem.

1. A generic protocol for application placement;
2. Instantiations for different management objectives.
3. The protocol can be run in a synchronous or asynchronous mode.

Three abstract methods:

1. `InitInstance()` is the initialization method for the specific gossip protocol.
2. `ChoosePeer()` is the method for selecting a peer for gossip interaction.



updatePlacement() is the method for recomputing the local state during a gossip interaction.

#### **4.2.SERVICE PROVIDER RESOURCE DEPLOYMENT**

Users access sites hosted by the cloud environment through the public Internet. A site is typically accessed through a URL that is translated to a network address through a global directory service, such as DNS. The components of the middleware layer run on all machines. The resources of the cloud are primarily consumed by module instances whereby the functionality of a site is made up of one or more modules. In the middleware, a module either contains part of the service logic of a site each machine runs a machine manager component that computes the resource allocation policy, which includes deciding the module instances to run. The resource allocation policy is computed by a protocol (later in the paper called P\*) that runs in the resource manager component. This component takes as input the estimated demand for each module that the machine runs. The computed allocation policy is sent to the Module scheduler for implementation/execution, as well as the site managers for making decisions on request forwarding.

#### **4.3. SERVICE OVERLAY MANAGEMENT**

The overlay manager implements VIRTUAL machines in the cloud and provides each APPLICATION list of machines to interact with. The overlay station approximate responsibility means in the large cloud environment n number of application and sites are running, each and every sites, application are running by virtual machine to maintained by the Large Cloud Environment, here what we do means to make the user graph of the specific application and the sites.

#### **5. ADAPT SERVICE MANAGEMENT**

The authorization hosted their sites/application in the Large Cloud Environment. The Service Level Agreement (SLA) and fine grained from the Cloud service provider and the Authorization. The Service level objectives from the authorization and the site user are also fine grained. In future the authorization needs the

enhance such requirement to the own sites and the application. They Designed such features to add their hosted. We address that the placing modules identically instance of modules on machine allocated in cloud resources.

#### **6. CONCLUSION**

We are implement a gossip protocol that computes in distributed and continuous fashion, a heuristic solution to resource allocation problem for dynamically changes the resource demand. We evaluate the protocol. we make a significant contribution towards engineering a resource management middleware for cloud environments. We identify a key component of such a middleware and present a protocol that can be used to meet our design goals for resource management: fairness of resource allocation with respect to sites, efficient adaptation to load changes and scalability of the middleware layer in terms of both the number of machines in the cloud as well as the number of hosted sites/applications. We presented a gossip protocol P\* that computes, in a distributed and continuous fashion, a heuristic solution to the resource allocation problem for a dynamically changing resource demand.

#### **7. REFERENCES**

- [1] M. Jelasity, A. Montresor, and O. Babaoglu, "Gossip-based aggregation in large dynamic networks," ACM Trans. Computer Syst., vol. 23, no. 3, pp. 219–252, 2005.
- [2] "T-Man: gossip-based fast overlay topology construction," Computer Networks, vol. 53, no. 13, pp. 2321–2339, 2009.
- [3] F. Wuhib, R. Stadler, and M. Spreitzer, "Gossip-based resource management for cloud environments," in 2010 International Conference on Network and Service Management.

# Analysis to Online Stock for Decision Approach of Investor

G.Magesh M.Sc., M.Phil, S. Saradha M.Sc., M.Phil,(Ph.d),,  
1,2,(Dept of Computer Science,, Vels University, Chennai)

\*\*\*\*\*

## Abstract:

The Internet service provides the wider opportunity for the investors to post online opinions that they share with fellow investors. Attitude analysis of online opinion posts can be facilitating both investors investment on decision making and stock companies risk perception. In this paper develops novel sentiment ontology to conduct context-sensitive sentiment analysis of stock markets are opinion post on online. A typical financial has been selected as an experimental platform of financial review data was collected. Computational results show that the statistical machine learning approach has higher classification perfection than the semantic approach. Then results also imply that investor sentiment has a particularly strong effect for rate of stocks relative to growth stocks. It has been reported that these message boards can have a significant impact reflect on the financial markets.

**Keywords — Accuracy, ontology, stock markets, financial and Sentiment**

\*\*\*\*\*

## I. Introduction

Investors have often found it very costly to acquire useful information to assist them in investment decision making. many investors have devoted a great deal of time to read messages posted on internet stock message boards to rate asset values based on information of varying quality. It has been reported that these message boards can have the significant impact on financial markets. Efficient investment decision making today is based on a different information sources that including historical financial data series and messages posted on stock message boards. There is an obvious correlation between investor sentiment and stock market performance. The investor sentiment predict future stock price. A sentiment index would be useful. Research on sentiment index selection has been identified both the direct sentiment and the indirect sentiment indices. However, a direct sentiment index based on questions and an indirect sentiment index based on related stock market data can be inaccurate. With the development of the Internet services, most of the investors have shared their opinions on stocks via stock forums, providing a great deal of discussion information. The Insider information and rumors can be an important

communication platform for investors. Stock information can be reflecting in investor sentiment and it can play an important role in investor decision making. These opinions could be alter the investing way of invertors in trade, acquire, and share information. Extracting investor sentiment and producing a sentiment index from the stock forums would be valuable. Overall, our study shows that stock forum sentiments does contain valuable information to decision making for investing and reinforce the investor sentiment speculation that irrational investors do influence stock markets.

## 1.2 Objective

Stock information that can reflect on investor sentiment and it can play the important role in decision making of investors. Those opinions could alter the way of the investing method of investors, acquire, and share information. Extracting investor sentiment and creating a sentiment index from stock forums would be very valuable.

## 2. Literature Survey

**Title:** Economic Forces, Emerging Eastern European Stock Markets and Sentiment

**Author:** Dmitrij Celov, Žana Grigaliuniene

**Year :** 2010

**Description** The aim of the current study is to explore the effects of macroeconomic news on stock returns to Eastern European countries, combining market and macroeconomic data over the period of 2000-09, during which that markets experienced excessive optimistic and pessimistic episodes. Hypothesizing the asymmetry in stock price responded to good and bad news, which seek to test its degree under the specific market conditions. The flaw correction models for each country are extended with fixed effects panel data specification, for capturing the cross-sectional effects of the state on the market to return responses to macroeconomic news. The key methodological to analysis problem in recent research is how to relate the daily stock data and monthly macroeconomic data. The aggregation of the stock returns to monthly averages have a several advantages over macroeconomic data disaggregation to irregular frequencies or calendar methods. The macroeconomic data is large1, witch exploit data mining techniques to judiciously fit the monthly averages of the stock returns and panel data analysis of capture the common patterns typical to Eastern European stock markets.

### **3. Project Description**

#### **3.1. General**

We analyze the problem of behavioral characterization of Stock information can reflect investor sentiment and can play the important role in investor decision making. These opinions could alter this way in which the investors invest trade, acquire, and share information.

#### **3.2. Problem Definition**

A direct sentiment index based on questions and an indirect sentiment index based on related stock market data can be inaccurate. With the development of the Internet service, more investors have shared their opinions on stocks via stock forums, providing a great deal of discussion information. Associate information and rumors can

be an important communication platform for investors.

### **3.3. Methodologies**

Stock information can reflect investor sentiment and can play an important role in investor decision making. Those opinions could alter this way in which investors invest trade, acquire, and share information. Extracting investor sentiment and introducing a sentiment index from stock forums would be valuable.

## **4. Modules**

### **❖ Investor**

- Authentication.
- View product details.

### **❖ User Opinion**

- Direct Index
- Indirect Index

### **❖ Admin**

- Stock analysis
- Company details.
- 

## **Module Description**

### **Authentication**

If you are the new user going to consume the service then they have to register on the first time implement for the necessary details. After successful completion of sign up process, the user has to login into the application by implement for username and exact password. The user has to provide exact username and password which is provided at the time of registration, if login success means it will take to master page else it will display the error message and it stay in login page itself.

### **View product detail**

The user has to provide exact username and password which is provided at the time of registration, if login success means it will take to master page else it will display the error message and it stay in login page itself. Investor search product details and view product details.

### **Direct Index**

An investor purchases a share of an index fund, he or she is purchasing a share of a portfolio that contains the good securities in an underlying index. The index fund holds the securities in this same proportion as they occur in the actual index and whenever the index decreases in value,

the fund's shares decrease as well, and vice versa. This only time an index buys or sells a stock is when the index itself. Index funds have ticker symbols and are traded on all major exchanges.

### **Indirect Index**

Indirect sentiment index based on related stock market data can be inaccurate. With a development of the Internet server most of investors have shared their opinions on stocks via stock forums, providing a great deal of discussion information.

### **Stock Analysis**

We use sentiment analysis technology to automatically classify unstructured reviews as positive or negative and they identify investor sentiment as either bullish or bearish.

### **company details**

Stock markets mean the transfer for money of a stock or security from a seller to the buyer. This requires these two parties to agree on a price. (Stocks or shares) confer an ownership interest in a particular company.

### **5. Conclusion**

I conducted three experimental scenarios: the comparison to the classification performance between a machine learning approach and a lexicon approach, the two experiments involving that relationship analysis of sentiment and stock price volatility learning. This to compare different methods, our results demonstrated that the statistical machine learning approach with a classification accuracy of 81.82%, which is higher than that of the semantic approach with a classification accuracy of 75.58%, significant at the 95% level. In this classification accuracy of the statistical machine learning approach was reasonably robust with respect to the size of the training set when the size was more than 600. To examine the relationship analysis of sentiment and the stock price volatility learning at the industry level, we varied the order of the sentiment related terms using parameters, such as order of the GARCH terms and the order of the ARCH terms in the GARCH-SVM models unchanged. Illustration

results suggested that improving the values of the order of the sentiment-related terms would benefit model [organization](#) accuracy when the order of the sentiment-related term was smaller than a threshold value. This model classification power would not benefit from changing the order of the sentiment-related terms when it exceeded the threshold value.

### **6. References**

- [1] B. Watkins, "Riding the wave of sentiment: An analysis of return consistency as a predictor of future returns," *J. Behav. Fin.*, vol. 4, no. 4, pp. 191–200, 2003.
- [2] D. Celov and Z. Grigaliūnas, "Economic forces, sentiment and emerging Eastern European stock markets," *Res. Econ. Bus. Central Eastern Eur.*, vol. 2, no. 2, pp. 37–53, 2010.
- [3] T. Lux, "Sentiment dynamics and stock returns: The case of the German stock market," *Empirical Econ.*, vol. 41, no. 3, pp. 663–679, 2011.
- [4] T. Zhang, J. Li, and P. Malone, "Closed-end fund discounts in Chinese stock markets," *Chin. Econ.*, vol. 37, no. 3, pp. 17–38, 2004.
- [5] Y. Wang and A. Di Iorio, "The cross section of expected stock returns in the Chinese A-share market," *Global Fin. J.*, vol. 17, no. 3, pp. 335–349, 2007.
- [6] H.-J. Sheu, Y.-C. Lu, and Y.-C. Wei, "Causalities between sentiment indicators and stock market returns under different market scenarios," *Int. J. Bus. Fin. Res.*, vol. 4, no. 1, pp. 159–171, 2010.
- [7] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *J. Fin.*, vol. 61, no. 4, pp. 1645–1680, 2006.

# Cluster Ensemble Approach for Clustering Mixed Data

Honorine Mutazinda A<sup>1</sup>, Mary Sowjanya<sup>2</sup>, O.Mrudula<sup>3</sup>

<sup>1,2,3</sup>( M.Tech, Department of Computer Science and Systems Engineering, Andhra University/College of Engineering, Visakhapatnam India)

\*\*\*\*\*

## Abstract:

The paper presents a clustering ensemble method based on ensemble clustering approach for mixed data. A clustering ensemble is a paradigm that seeks to best combine the outputs of several clustering algorithms with a decision fusion function to achieve a more accurate and stable final output.

Most traditional clustering algorithms are limited to handling datasets that contain either numeric or categorical attribute and these algorithms were not generally scalable for large datasets. However; datasets with mixed types of attributes are common in real life data mining applications. So a novel divide-and-conquer technique is designed and implemented to solve this problem.

First, the original mixed dataset is divided into two sub-datasets: the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of datasets are employed to produce corresponding clusters. Last, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which the categorical data clustering algorithm is used to get the final clusters.

**Keywords — Clustering, Novel divide-and-conquer, Mixed Dataset, Numerical Data, and Categorical Data.**

\*\*\*\*\*

## I. Introduction

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining models (prediction and description) are achieved by using the following primary data mining tasks: Classification, Regression, Clustering, Summarization, and Dependency modelling and Change and Deviation Detection. Clustering groups elements in a data set in accordance with its similarity such that elements in each cluster are similar while elements from different clusters are dissimilar. It involves analyzing or processing multivariate data, such as: characterizing customer groups based on purchasing patterns, categorizing Web documents, grouping genes and proteins that have similar functionality, grouping spatial locations prone to earthquakes based on seismological data, etc. Clustering ensembles or clustering fusion is the integration of results from

various clustering algorithms using a consensus function to yield stable results. The idea of combining different clustering results (cluster ensemble or cluster aggregation) emerged as an alternative approach for improving the quality of the results of clustering algorithms.

In this paper a cluster ensemble approach using divide and conquer technique has been designed and implemented to deal with such type of mixed datasets. So, the initial dataset is divided into sub datasets namely numerical and categorical. Then clustering algorithms designed for numerical and categorical datasets can be employed to produce corresponding clusters. Finally, the clustering results from the above step are combined as a categorical dataset on which the same categorical clustering algorithm or any other can be used to produce the final output clusters.

## II. Related Work



### **A. Clustering Mixed Data**

Most of the traditional clustering algorithms are designed to focus either on numeric data or on categorical data. The collected data in real world often contain both numeric and categorical attributes. It is difficult for applying traditional clustering algorithm directly into these kinds of data. Typically, when people need to apply traditional distance-based clustering algorithms to group these types of data, a numeric value will be assigned to each category in this attributes. Some categorical values, for example “low”, “medium” and “high”, can easily be transferred into numeric values. But if categorical attributes contain the values like “red”, “white” and “blue” ... etc., it cannot be ordered naturally.

Due to the differences in their features, in order to group these assorted data, it is good to exploit the clustering ensemble method which uses split and merge approach to solve this problem. For clustering mixed type attributes in [1] Ming-Yi Shih presented a new two-step clustering method is presented to find clusters on Mixed Categorical and Numeric Data. Items in categorical attributes are processed to construct the similarity or relationships among them based on the ideas of co-occurrence; then all categorical attributes can be converted into numeric attributes based on these constructed relationships. Finally, since all categorical data are converted into numeric, the existing clustering algorithms can be applied to the dataset without pain.

Jongwoo Lim<sup>1</sup>, Jongeun Jun<sup>2</sup> in [2] proposed a clustering framework that supports clustering of datasets with mixed attribute type (numerical, categorical), while minimizing information loss during clustering. They first utilize an entropy based measure of categorical attributes as a criterion function for similarity. Second, based on the results of entropy based similarity, they extract candidate cluster numbers and verify their weighting scheme with pre-clustering results. Finally, they cluster the mixed attribute type datasets with the extracted candidate cluster numbers and the weights.

Zhexue Huang in [6] presented a k-prototypes algorithm which is based on the k-means

paradigm but removes the numeric data limitation whilst preserving its efficiency. In the algorithm, objects are clustered against k prototypes. A method is developed to dynamically update the k prototypes in order to maximize the intra cluster similarity of objects. When applied to numeric data the algorithm is identical to the k-means. To assist interpretation of clusters we use decision tree induction algorithms to create rules for clusters. These rules, together with other statistics about clusters, can assist data miners to understand and identify interesting clusters.

Jamil Al-Shaqsi and Wenjia Wang in [7] present a clustering ensemble method based on a novel three-staged clustering algorithm. Their ensemble is constructed with a proposed clustering algorithm as a core modeling method that is used to generate a series of clustering results with different conditions for a given dataset. Then, a decision aggregation mechanism such as voting is employed to find a combined partition of the different clusters. The voting mechanism considered only experimental results that produce intra-similarity value higher than the average intra-similarity value for a particular interval. The aim of this procedure is to find a clustering result that minimizes the number of disagreements between different clustering results.

### **B. Cluster Ensemble**

Clustering fusion is the integration of results from various clustering algorithms using a consensus function to yield stable results.

The idea of combining different clustering results (cluster ensemble or cluster aggregation) emerged as an alternative approach for improving the quality of the results of clustering algorithms. It is based on the success of the combination of supervised classifiers. Given a set of objects, a cluster ensemble method consists of two principal steps: Generation, which is about the creation of a set of partitions of these objects, and Consensus Function, where a new partition, which is the integration of all partitions obtained in the generation step, is computed.

Generation Mechanisms:

Generation is the first step in clustering ensemble methods, in which the set of clusterings is generated and combined. It generates a collection of clustering solutions i.e., a cluster ensemble. Given a data set of  $n$  instances  $X = \{X_1, X_2, \dots, X_n\}$  an ensemble constructor generates a cluster ensemble, represented as  $\Pi = \{\pi^1, \dots, \pi^r\}$  where  $r$  is the ensemble size (the number of clustering in the ensemble).

Each clustering solution  $\pi^i$  is simply a partition of the data set  $X$  into  $K_i$  disjoint clusters of instances, represented as  $\pi^i = C_1^i, \dots, C_{K_i}^i$ .

It is very important to apply an appropriate generation process, because the final result will be conditioned by the initial clusterings obtained in this step.

In the generation step there are no constraints about how the partitions must be obtained. Therefore, in the generation process different clustering algorithms or the same algorithm with different parameters initialization can be applied.

1) **Consensus Functions:** The consensus function is the main step in any clustering ensemble algorithm. In this step, the final data partition or consensus partition  $P^*$ , which is the result of any clustering ensemble algorithm, is obtained. However, the consensus among a set of clusterings is not obtained in the same way in all cases. There are two main consensus function approaches: objects co-occurrence and median partition.

### III. A Cluster Ensemble Approach for

#### Clustering Mixed Data

##### C. Overview

In This approach, instead of  $k$  means that assumes clusters are hyper-ellipsoidal and of similar sizes and which can't find clusters that vary in size to cluster numerical dataset, Chameleon an agglomerative hierarchical algorithm is chosen. This is due to the fact that Chameleon considers the

internal characteristics of the clusters and can automatically adapt to the merged clusters. Also it can better model the degree of interconnectivity and closeness between each pair of clusters than  $K$  means. The existing Squeezer algorithm to cluster categorical data is retained as it is suitable for handling data streams and also can handle outliers effectively.

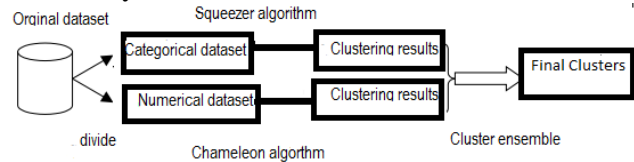


Figure 1. Overview of algorithm framework.

##### Algorithm

1. Splitting of the given data set into two parts. One for numerical data and another for categorical data.
2. Applying clustering Chameleon algorithms for numerical data set.
3. Applying clustering Squeezer algorithms for categorical data set.
4. Applying clustering Squeezer algorithms for categorical data set
5. Combining the output of step 2 and step 3 as cluster ensemble
6. Clustering the results using squeezer algorithm.
- 3) Final resultant clusters. Chameleon algorithm: Chameleon is a new agglomerative hierarchical clustering algorithm that overcomes the limitations of existing clustering algorithms. The Chameleon algorithm's key feature is that it accounts for both interconnectivity and closeness in identifying the most similar pair of clusters.

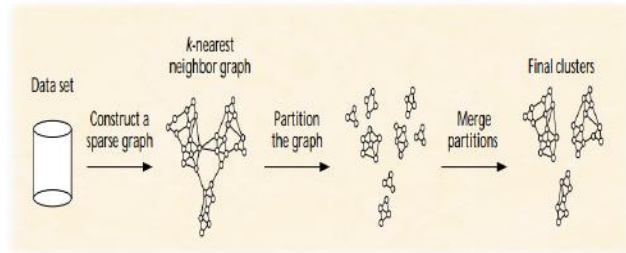


Figure 2. Overview of chameleon algorithm

Chameleon uses a two-phase algorithm, which first partitions the data items into sub-clusters and then repeatedly combines these sub-clusters to obtain the final clusters. It first clusters the data items into several sub-clusters that contain a sufficient number of items to allow dynamic modelling.

Chameleon uses a dynamic modelling framework to determine the similarity between pairs of clusters by looking at their relative interconnectivity (RI) and relative closeness (RC). Chameleon selects pairs to merge for which both RI and RC are high. That is, it selects clusters that are well interconnected as well as close together.

- **Relative interconnectivity:** Clustering algorithms typically measure the absolute interconnectivity between clusters  $C_i$  and  $C_j$  in terms of edge cut—the sum of the weight of the edges that straddle the two clusters, which we denote  $EC(C_i, C_j)$ .

Relative interconnectivity between clusters is their absolute interconnectivity normalized with respect to their internal interconnectivities. To get the cluster's internal interconnectivity, we sum the edges crossing a min-cut bisection that splits the cluster into two roughly equal parts.

Thus, the relative interconnectivity between a pair of clusters  $C_i$  and  $C_j$  is:

$$RI(C_i, C_j) = \frac{|EC(C_i, C_j)|}{\frac{|EC(C_i)| + |EC(C_j)|}{2}}$$

- **Relative closeness:** Relative closeness involves concepts that are analogous to

those developed for relative interconnectivity.

The absolute closeness of clusters is the average weight (as opposed to the sum of weights for interconnectivity) of the edges that connect vertices in  $C_i$  to those in  $C_j$ .

To get a cluster's internal closeness, we take the average of the edge weights across a min-cut bisection that splits the cluster into two roughly equal parts. The relative closeness between a pair of clusters is the absolute closeness normalized with respect to the internal closeness of the two clusters:

$$RC(C_i, C_j) = \frac{\overline{SEC}(C_i, C_j)}{\frac{|C_i|}{|C_i| + |C_j|} \overline{SEC}(C_i) + \frac{|C_j|}{|C_i| + |C_j|} \overline{SEC}(C_j)}$$

Where  $\overline{SEC}(C_i)$  and  $\overline{SEC}(C_j)$  are the average weights of the edges that belong in the min-cut bisector of clusters  $C_i$  and  $C_j$ , and  $SEC(C_i, C_j)$  is the average weight of the edges that connect vertices in  $C_i$  and  $C_j$ . Terms  $|C_i|$  and  $|C_j|$  are the number of data points in each cluster.

#### ➤ Advantages

Existing clustering algorithms find clusters that fit some static model. Although effective in some cases, these algorithms can break down—that is, cluster the data incorrectly.

They break down when the data contains clusters of diverse shapes, densities, and sizes. Existing algorithms use a static model of the clusters and do not use information about the nature of individual clusters as they are merged.

#### 4) Squeezer Algorithm:

The Squeezer algorithm has  $n$  tuples as input and produces clusters as final results. Initially, the first tuple in the database is read in and a Cluster Structure (CS) is constructed with  $C = \{1\}$ . Then, the subsequent tuples are read iteratively. For each tuple, by our similarity function, we compute its similarities with all existing clusters, which are represented and embodied in the corresponding CSs. The largest value of similarity is selected out. If it is larger than the

given threshold, denoted as  $s$ , the tuple is put into the cluster that has the largest value of similarity. The CS is also updated with the new tuple. If the above condition does not hold, a new cluster must be created with this tuple. The algorithm continues until all tuples in the dataset are traversed.

The sub-function `addNewClusterStructure()` uses the new tuple to initialize Cluster and Summary, and then a new CS is created. The sub-function `addTupleToCluster()` updates the specified CS with new tuple. The subfunction `simComputation()`, which makes use of information stored in the CS to get the statistics based similarity.

**Algorithm Squeezer(D,s) sim**

Begin

1. While(D has unread tuple) {
2. Tuple=get Current Tuple(D)
3. If (tuple.tid==1) {
4. AddNewClusterStructure (tuple.tid)}
5. else {
6. for each existing cluster C
7. SimComputation(C,tuple)
8. get the max value of similarity:  $sim\_max$
9. Get the corresponding Cluster Index: index
10. if  $sim\_max \geq s$
11. addTupleToCluster(tuple,index)
12. else
13. addNewClusterStructure(tuple.tid)}
14. } handle outliers()
15. output ClusteringResults()

End

➤ **Advantages**

- The Squeezer algorithm only makes one scan over the dataset, thus, is highly efficient for disk resident datasets where the I/O cost becomes the bottleneck of efficiency.
- The algorithm is suitable for clustering data streams, where given a sequence of points, the objective is to maintain consistently good clustering of the sequence so far, using a small amount of memory and time.
- Outliers can be handled efficiently and directly.
- The algorithm does not require the number of desired clusters as an input parameter. This is very important for the user who usually does not know this number in advance.

## **IV. Methodology**

A divide and conquer approach for mixed data using cluster ensemble works effectively either on pure numeric data or on pure categorical data.

In this approach first, the original mixed dataset is divided into two sub-datasets: the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of datasets can be employed to produce corresponding clusters. Here Chameleon algorithm is used for the numeric dataset where as Squeezer is used for the categorical dataset. In the Last step, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which any categorical data clustering algorithm can be used to get the final clusters and Squeezer which was previously used for categorical data is again used for this purpose.

**Algorithm**

**Step1.** Splitting of the given data set into two parts. One for numerical data and another for categorical data.

**Step 2.** Applying clustering chameleon algorithms for numerical data set

**Step 3.** Applying clustering squeezer algorithms for categorical data set

**Step 4.** Combining the output of step 2 and step 3

**Step 5.** Clustering the results using squeezer algorithm.

**Step6.** Final cluster results.

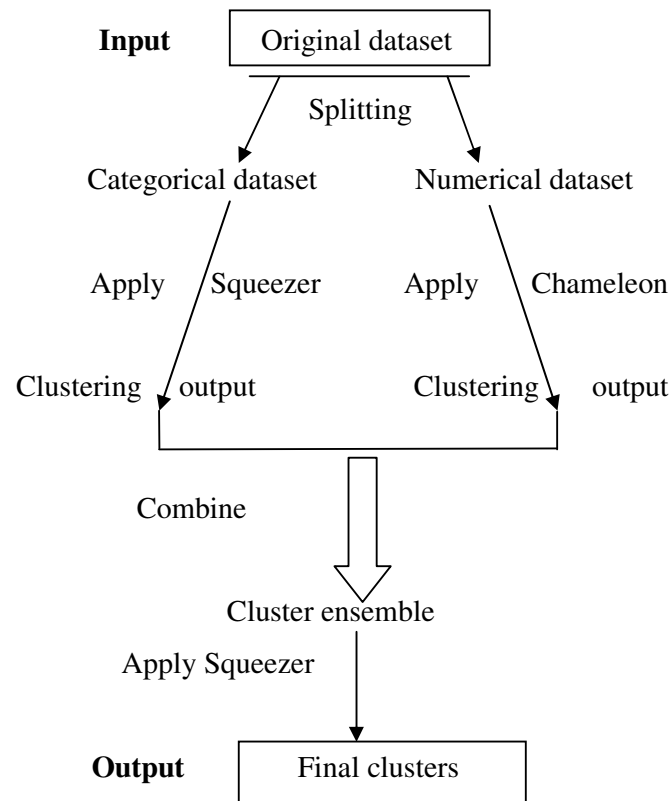


Figure 3. Algorithm for proposed framework

## V. Experimental Results

### D. Adult Dataset

```

30. Self-emp-not-inc, 393606, Some-college, 10, Divorced, Sales, Unmarried, White, Male, 0.0.65, United-States'
31. Private, 39150, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0.0.40, United-States'
32. Private, 288840, HS-grad, 9, Married-spouse-absent, Other-service, Unmarried, Black, Female, 0.0.38, United-States'
34. Private, 232703, Some-college, 10, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0.0.40, Philippines'
42. Private, 79586, Bachelors, 13, Married-civ-spouse, Adm-clerical, Husband, Asian-Pac-Islander, Male, 0.0.40, United-States'
48. Self-emp-not-inc, 82098, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Asian-Pac-Islander, Male, 0.0.65, United-States'
38. Private, 245722, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, White, Male, 7088.0.45, United-States'
29. Private, 78261, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0.0.40, United-States'
33. Private, 355396, 10th, 6, Never-married, Handlers-cleaners, Not-in-family, White, Male, 0.0.40, United-States'
54. Private, 218490, Bachelors, 13, Divorced, Exec-managerial, Not-in-family, White, Male, 27020.0.55, United-States'
44. Private, 110908, Assoc-voc, 11, Married-civ-spouse, Transport-moving, Wife, White, Female, 0.0.25, United-States'
42. Federal-gov, 34218, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 7298.0.50, United-States'
49. Private, 248095, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0.0.45, United-States'
25. Private, 363707, HS-grad, 9, Married-civ-spouse, Handlers-cleaners, Husband, White, Male, 0.0.40, United-States'
33. Private, 272411, Bachelors, 13, Never-married, Exec-managerial, Not-in-family, White, Female, 0.0.40, United-States'
38. Private, 120033, Some-college, 10, Married-civ-spouse, Sales, Husband, White, Male, 0.0.60, United-States'
20. Private, 172287, HS-grad, 9, Never-married, Sales, Not-in-family, White, Female, 0.0.38, United-States'
44. Private, 177344, HS-grad, 9, Never-married, Handlers-cleaners, Not-in-family, White, Male, 0.0.50, United-States'
45. Private, 235458, Bachelors, 13, Married-civ-spouse, Transport-moving, Husband, White, Male, 0.0.40, United-States'
27. Self-emp-inc, 193868, HS-grad, 9, Never-married, Sales, Not-in-family, White, Male, 0.0.50, United-States'
18. Private, 232082, HS-grad, 9, Never-married, Machine-op-inspct, Own-child, White, Male, 0.0.40, United-States'
38. Private, 27400, HS-grad, 9, Divorced, Craft-repair, Unmarried, White, Male, 0.0.50, United-States'
45. Private, 247043, 11th, 7, Married-civ-spouse, Craft-repair, Husband, White, Male, 0.0.42, United-States'
27. Local-gov, 162404, HS-grad, 9, Never-married, Protective-serv, Not-in-family, Black, Male, 2174.0.40, United-States'
64. Private, 236341, 5th-6th, 3, Widowed, Other-service, Not-in-family, Black, Female, 0.0.16, United-States'
66. Local-gov, 179285, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 3432.0.20, United-States'
34. Private, 30433, Bachelors, 13, Never-married, Sales, Not-in-family, White, Female, 0.0.35, United-States'
42. Self-emp-not-inc, 102721, Bachelors, 13, Married-civ-spouse, Transport-moving, Husband, White, Male, 0.0.40, United-States'
26. Private, 375499, 10th, 6, Never-married, Adm-clerical, Not-in-family, Black, Male, 0.0.20, United-States'
27. Private, 178688, Assoc-voc, 11, Never-married, Craft-repair, Other-relative, White, Male, 0.0.40, United-States'
32. Private, 276709, Some-college, 10, Never-married, Sales, Other-relative, White, Female, 0.0.40, United-States'
23. 7, 438087, Some-college, 10, Never-married, ?, Own-child, White, Male, 0.0.30, United-States'
47. Private, 84790, Some-college, 10, Married-civ-spouse, Craft-repair, Husband, White, Male, 0.0.40, United-States'
20. State-gov, 37482, Some-college, 10, Never-married, Adm-clerical, Own-child, White, Female, 0.0.40, United-States'
46. State-gov, 178686, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0.0.38, United-States'
35. 7, 153926, HS-grad, 9, Married-civ-spouse, ?, Wife, Black, Female, 0.0.40, United-States'
52. Private, 110740, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0.1897.40, United-States'
28. Private, 116613, Some-college, 10, Never-married, Tech-support, Own-child, White, Female, 0.0.24, United-States'
21. Private, 108687, Some-college, 10, Never-married, Sales, Own-child, White, Female, 0.0.40, United-States'
26. Private, 355739, HS-grad, 9, Never-married, Machine-op-inspct, Not-in-family, White, Male, 0.0.40, United-States'
29. Private, 195284, Doctorate, 16, Divorced, Prof-specialty, Not-in-family, White, Female, 0.0.60, United-States'
38. Private, 125333, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0.0.40, ?
37. Private, 140054, Bachelors, 13, Marr, Craft-repair, Husband, White, Male, 1808.734025.06.305951.40.426824.7
  
```

C:\Users\NWP\workspace9\EnsembleClustering\src>



Figure 4. Adult dataset

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
5797	1788	4121	3723	2189	8254	3722
42.8039	40.6158	26.6518	46.2812	36.1969	43.9701	25.9422
186894.2053	180243.4754	185121.238	182868.2893	208583.688	188321.8828	203562.9565
13.1832	10.16	19.3028	9.4225	9.8305	8.6886	9.817
3254.8357	1422.7819	251.0806	443.3932	471.0539	777.1381	181.8044
158.1172	97.57	51.0702	47.5326	65.2197	94.3276	37.4234
44.9997	36.722	32.5547	38.1034	41.2284	43.4895	38.8611
Private	Private	Private	Private	Private	Private	Private
Bachelors	HS-grad	Some-college	HS-grad	HS-grad	HS-grad	HS-grad
Married-civ-spouse	Married-civ-spouse	Never-married	Divorced	Never-married	Married-civ-spouse	Never-married
Prof-specialty	Adm-clerical	Other-service	Other-service	Adm-clerical	Craft-repair	Craft-repair
Husband	Wife	Own-child	Unmarried	Not-in-family	Husband	Own-child
White	White	White	White	Black	White	White
Male	Female	Female	Female	Male	Male	Male
United-States	United-States	United-States	United-States	United-States	United-States	United-States

Figure 5. Clustering mixed data with k=7 on Adult dataset

### E. Credit approval dataset

The credit approval dataset has 690 instances, each being described by 6 numeric and 9 categorical attributes. Instances were classified into two classes, "+" for approved label and "-" for rejected label.

The comparison has been done with k-prototype algorithm that has been applied on credit approval dataset previously [10]. Clustering accuracy to measure clustering results is as follows:

$$r = \frac{\sum_{i=1}^k a_i}{n}$$

```

b.27.42.12.5.u.g.ss.bb.0.25.f.f.0.t.g.720.0.Rejected
b.24.75.0.54.u.g.m.v.1.f.f.0.t.g.120.1.Rejected
b.41.17.1.25.v.p.v.v.0.25.f.f.0.f.g.0.125.Rejected
a.33.008.1.625.u.g.d.v.0.54.f.f.0.t.g.0.0.Rejected
b.27.83.2.04.v.p.x.h.0.04.f.f.0.f.g.120.1.Rejected
a.23.58.0.585.v.p.ff.ff.0.125.f.f.0.f.g.120.37.Rejected
b.26.17.12.5.v.p.k.h.1.25.f.f.0.t.g.0.12.Rejected
b.31.2.005.u.g.c.v.0.005.f.f.0.f.g.300.0.Rejected
b.20.75.5.005.v.p.j.v.0.29.f.f.0.f.g.140.184.Rejected
b.28.92.0.375.u.g.c.v.0.29.f.f.0.f.g.220.140.Rejected
a.51.92.6.5.u.g.i.bb.3.005.f.f.0.t.g.70.0.Rejected
a.22.67.0.335.u.g.v.0.75.f.f.0.f.s.160.0.Rejected
b.34.5.005.v.p.i.bb.1.005.f.f.0.t.g.400.0.Rejected
a.69.5.6.u.g.ff.ff.0.f.f.0.f.s.0.0.Rejected
a.40.33.8.125.v.p.k.v.0.165.f.t.2.f.g.7.18.Rejected
a.19.50.0.665.v.p.c.v.1.f.t.1.f.g.2000.2.Rejected
b.16.3.125.u.g.v.v.0.005.f.t.1.f.g.0.6.Rejected
b.17.008.0.25.u.g.v.v.0.335.f.t.4.f.g.160.8.Rejected
b.31.25.2.835.u.g.ff.ff.0.f.t.5.f.g.176.146.Rejected
b.25.17.3.u.g.c.v.1.25.f.t.1.f.g.0.22.Rejected
a.22.67.0.79.u.g.i.v.0.005.f.f.0.f.g.144.9.Rejected
b.40.50.1.5.u.g.i.bb.0.f.f.0.f.s.300.0.Rejected
b.22.25.0.46.u.g.k.v.0.125.f.f.0.t.g.280.55.Rejected
a.22.25.1.25.v.p.ff.ff.3.25.f.f.0.f.g.280.0.Rejected
b.22.5.0.125.v.p.k.v.0.125.f.f.0.f.g.200.70.Rejected
b.23.50.1.72.u.g.c.v.0.54.f.f.0.t.g.136.1.Rejected
b.38.42.0.705.u.g.c.v.0.375.f.t.2.f.g.225.500.Rejected
a.26.58.2.54.v.p.ff.ff.0.f.f.0.t.g.180.60.Rejected
b.35.2.5.u.g.v.1.f.f.0.t.g.210.0.Rejected
b.20.42.1.005.u.g.v.1.5.f.f.0.f.g.100.7.Rejected
b.27.42.1.25.u.g.v.1.75.f.f.0.f.g.200.0.Rejected
b.26.17.0.835.u.g.c.v.1.165.f.f.0.f.g.100.0.Rejected
b.33.67.2.165.u.g.c.v.1.5.f.f.0.f.p.120.0.Rejected
b.24.50.1.25.u.g.c.v.0.25.f.f.0.f.g.110.0.Rejected
a.27.67.2.04.u.g.v.v.0.25.f.f.0.t.g.180.50.Rejected
b.37.5.0.835.u.g.e.v.0.84.f.f.0.f.g.120.5.Rejected
b.49.17.2.29.u.g.ff.ff.0.25.f.f.0.f.g.200.2.Rejected
b.33.58.0.335.v.p.c.v.0.005.f.f.0.f.g.180.0.Rejected
b.51.93.3.8.v.p.ff.ff.1.5.f.f.0.f.g.180.4.Rejected
b.22.92.3.165.v.p.c.v.0.165.f.f.0.f.g.160.1058.Rejected
b.21.83.1.54.u.g.k.v.0.005.f.f.0.t.g.356.0.Rejected
b.25.25.1.u.g.ss.v.0.5.f.f.0.f.g.200.0.Rejected
b.58.58.2.71.u.g.c.v.2.415.f.f.0.t.g.320.0.Rejected
b.19.50.0.585.u.g.ff.ff.0.f.t.3.f.g.350.769.Rejected
a.53.33.0.165.u.g.ff.ff.0.f.f.0.t.s.62.27.Rejected
a.22.17.1.25.u.g.ff.ff.0.f.t.1.f.g.92.300.Rejected

```

Figure 6. Credit approval dataset

*****Clustering results*****							
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
101	111	83	120	102	57	64	52
25.2284	28.768	40.6372	29.5633	34.3602	25.8598	36.4654	34.7634
2.8336	3.7714	8.4742	5.223	5.3167	4.9656	3.8423	3.367
0.9216	1.5561	5.7018	2.399	2.8023	0.8917	0.6528	2.2227
0.5842	0.4054	7.0482	4.7583	3.0294	0.5263	0.8281	0.0769
191.1785	197.8021	140.8072	184.3658	147.3339	165.0178	165.6411	324.2118
197.1586	1111.2252	2342.2892	1690.0917	1194.2157	193.1228	238.4219	258.4088
h	h	h	h	h	a	h	h
u	y	u	u	u	u	u	u
g	p	g	g	g	g	g	g
c	c	cc	q	c	i	ff	d
v	v	h	v	v	v	ff	v
f	f	t	t	f	f	f	f
t	t	t	t	f	f	f	t
g	g	g	g	g	g	g	g
Rejected	Rejected	Approved	Approved	Approved	Rejected	Rejected	Rejected

Figure 7. Clustering mixed data with k=8 on Credit Approval dataset

where n is the number of instances in the dataset,  $a_1$  is the number of instances occurring in both cluster I and its corresponding class, which has the maximum value.

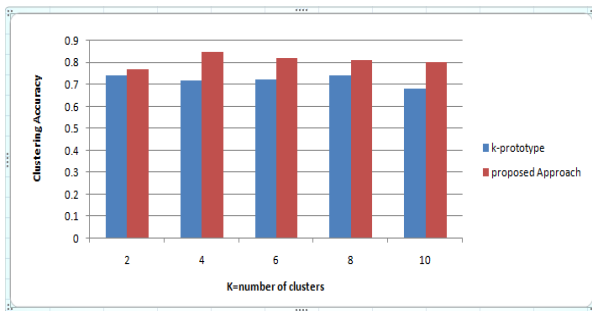


Figure 8. Clustering accuracy vs. number of clusters

## VI. Conclusions

Cluster analysis has been widely a tool to various domains to discover the hidden and useful patterns inside datasets. Previous clustering algorithms mostly focus on either numeric or categorical, recently, approaches for clustering mixed attribute type datasets have been emerged, but they are mainly based on transforming categorical to numerical attributes. Such approaches have disadvantages of poor results due to the loss of information because important portion of attribute values can be dropped out while transformation. Therefore, the proposed framework approach divides the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of

datasets can be employed to produce corresponding clusters. Here Chameleon algorithm is used for the numeric dataset where as Squeezer is used for the categorical dataset. Clustering results on the categorical and numeric dataset are combined as a categorical dataset, which allows integration of other algorithms to produce corresponding clusters which leads to a better clustering accuracy.

In future work, other clustering algorithms for large scale dataset with mixed attribute types can be explored, also some weighting schemes on existing algorithms to perform well on their corresponding type of attributes to improve the proposed framework.

## VII. References

- [1] Ming-Yi Shih\*, Jar-Wen Jheng and Lien-Fu Lai "A Two-Step Method for Clustering Mixed Categorical and Numeric Data" (2010).
- [2] Jongwoo Lim , Jongeun Jun , Seon Ho Kim and Dennis McLeod "A Framework for Clustering Mixed Attribute Type Datasets".
- [3] M. V. Jagannatha Reddy<sup>1</sup> and B. Kavitha" Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method "(2012).
- [4] George Karypis Eui-Hong (Sam) Han Vipin Kumar"Chameleon:Hierarchical Clustering Using Dynamic Modeling".

- [5] S. Anitha Elavarasi<sup>1</sup> and J. Akilandeswari<sup>2</sup> “survey on clustering algorithm and similarity measure for categorical data ”(2014).
- [6] zhexue huang “clustering large data sets with mixed numeric and categorical values”.
- [7] Jamil Al-Shaqsi and Wenjia Wang “A Clustering Ensemble Method for Clustering Mixed Data ”.
- [8] Zengyou He , Shengchun Deng ”Squeezer: An efficient algorithm for clustering categorical data”.
- [9] Dileep Kumar Murala “Divide and Conquer Method for Clustering Mixed Numerical and Categorical Data”(2013).
- [10] Zengyou He, Xiaofei Xu, Shengchun Deng “Clustering Mixed Numeric and categorical Data: A cluster Ensemble Approach”
- [11] divya d j<sup>1</sup> & gayathri devi b<sup>2</sup> “A meta clustering approach for ensemble problem ”.
- [12] Strehl and J. Ghosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.
- [13] G. Kharypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel hypergraph partitioning: Applications in vlsi domain. In *Proceedings of the Design and Automation Conference*, 1997.
- [14] Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In *Proceedings IEEE International Conference on Data Mining*, 2003.
- [15] M. Meila. Comparing clusterings by the variation of information. In *Proceedings of the Sixteenth Annual Conference of Computational Learning Theory (COLT)*, 2003.

# An Improved Collaborative Filtering Algorithm Based on Tags and User Ratings

CaiyunGuo<sup>1</sup>, HuijinWang<sup>2</sup>

<sup>1,2</sup>(College of Information Science and Technology, Jinan University, Guang dong, China)

\*\*\*\*\*

## Abstract:

Aiming at the problem that the existing social tags recommendation system in building user interest model does not fully reflect the genuine interests, this paper proposes aim proved recommended algorithm (TARBCF) based on tags and user ratings. Since the rating data often sparse, to make the best use of the both advantages of ratings and tags, a rating predicts algorithm based on item category is introduced to predict the ratings. In this paper, user's ratings can be incorporated to calculate the weight of tags. Considering the user interest has the time characteristic, time window is used to capture the current interests of user. Thus, by analyzing the traditional collaborative filtering thought, considering the relationship between user ratings and tags as well as the influence of user's current interest, this paper set up an user-tag correlation matrix, which can calculate the target user's nearest neighbors. Then according to the neighbor users predict the target user's preferences of candidate items. Finally, taking the top-N scores items recommend to the target user. Simulation experimental results show that the improved algorithm can better reflect the user's preferences, and the quality of its recommendations were superior to the traditional scheme.

**Keywords** —Collaborative filtering algorithm, Tags, Time window, User ratings

\*\*\*\*\*

## I. INTRODUCTION

With the huge amount of information is shared through network platforms, it is becoming more difficult for people to find the useful information in timely. Fortunately, Recommender System as an effective tool to solve the problem of "information overload" [1], which has been widely applied to various types of network platforms and e-commerce platforms(e.g., Amazon, eBay, and Taobao) [2].It is changing the way of people find information, which take the initiative to provide valuable information for the user. At present, how to build a good personalized recommendation system is still the focus of researchers, although there are a lot of works has been done in recent years.

Obviously, recommendation algorithm as the core part of personalized recommendation system, its accuracy closely related to the quality of recommendation system. Over the years, various approaches for building recommender systems have been developed that utilize either demographic, content, or historical information [3].Among them, collaborative filtering (CF) is the most successful technique in recommendation systems[4][5],it is a kind of recommendation algorithms that based on user's behavior data .The basic idea of CF is that if users have the similar behaviors in some items, they will rate or act on other items similarly, so it can through the neighbor users to decide whether to recommend the items to the target user.Compared with Content-based Recommendations (CB),CF is

more suitable for recommend non-text items, such as music, movies, pictures, etc., that is also one of the reasons why it has been widely used. However, with the network and user scale expands unceasingly, item quantity increased dramatically, sparse matrix and cold start problem become more and more serious, the recommend quality of the collaborative filtering was greatly reduced.

To alleviate the impact of data sparsity in collaborative recommendation system, a series of improved methods have been put forward by many researchers. An rating prediction scheme is applied to the typical CF algorithms, which is confirmed effectively to alleviate the data sparsity [6]. Singular value decomposition (SVD) model also be proposed to reduce the space dimension of matrix. This method can significantly improve the system expansion ability, but the dimension reduction will lead to information loss [7]. Considering the user's interest would be migrated over time, More and more researchers take time factor into consideration, which make the similarity calculation more accurate and a better recommendation result have been achieved [8].

These methods above are based on the rating data to improve the quality of recommendations. With the rapid development of the web2.0, social tag is widely applied to the recommendation system, users can choose or mark tags according to their own understanding and preferences, which contains a lot of user interests information. Thus, taking tags into recommendation algorithm can help us to improve the quality of the recommendations. There are several studies that exploit various aspects of tags to build user interest modeling. Au Yeung, Cibbins, and Shadbolt constructed a user's model which can represent multiple interests of the user by forming a set of frequent tag patterns [9]. Nakamoto al. proposed Reasonable Tag-based CF(RCF) that first clusters tags into topics by using

an expectation-maximization (EM) algorithm [10]. More recently, Wang, Clements, and Reinders introduced a collaborative tagging model, a collaborative browsing model, and a collaborative item search model into building personalization models [11]. Thus, the usage of tags allows us to capture valuable information for understanding user interests and can build better interest models.

Thus, on the basis of these studies, inspired by the idea of combining rating information with tags, an improved collaborative filtering algorithm based on tags and user ratings is proposed in this paper, which is improved through two ways: One is improving user similarity calculating method, and the other is catching the target user interest in recommendation generation phase.

## II. DEFINITIONS

To better describe the proposed method in this paper, we define some key concepts and used in this paper as below:

- Users:  $U = \{u_1, u_2, \dots, u_{|U|}\}$  contains all users who have used tags or rating to evaluate items.
- Items (i.e., Resources):  $I = \{i_1, i_2, \dots, i_{|I|}\}$  contains all items evaluated by users in  $U$ . It could be any type of resource or products that come from our daily life, such as videos, music, movies etc.
- Tags:  $T = \{t_1, t_2, \dots, t_{|T|}\}$  contains all tags used by users in  $U$ . A tag is a piece of textual information and can be used by users to label multiple items.
- Item Category:  $C = C_1 \cup C_2 \cup \dots \cup C_k$  contains all item category information in the system. An item may belong to several categories,



where  $C_j = \{I_{j,1}, I_{j,2}, \dots, I_{j,k}\}$  represents the item set that belong to the  $j$ -th category.

- User-Item Rating Matrix: Described as a  $|U| \times |I|$  ratings matrix  $R_{|U| \times |I|} = (R_{i,j})_{|U| \times |I|}$ . The row represents  $|U|$  users and column represents  $|I|$  items. The element of matrix  $R_{i,j}$  means the rating rated to the user  $i$  on the item  $j$ , which is acquired with the rate of user's interest. Usually, the ratings is on a 1-5 scale or unknown.

### III. THE PROPOSED APPROACHES

As we all know, tags are popularly used in various kinds of application areas, it is becoming another important implicit rating information used to profile users' interests. However, the tags used by users are free-formed and contain semantic ambiguities and tag synonyms [12].

In this paper, we proposed a new method to integrate tags and ratings to improve the accuracy of predictions based on the traditional CF. Different from the earlier work, we focus on what a tagged item is about and how much a user prefers the item, rather than capturing what tags are used by the user.

Since the explicit ratings are rare or not available in real life, we should use the rating predict algorithm to predict the rating of no rating items. Then, using those ratings and tags information to build User-Tag Correlation Matrix, which can integrate the advantages of ratings and tags. Finally, according the neighbors of target user, we can make recommendation by using the similarity information of items. The specific steps are as follows:

#### A. Rating Predict Algorithm Based on Item Category

First, classifying the items by different categories, then using the Item-based collaborative filtering algorithm (IBCF) to predict the rating of the no rating items in every category. Since an item may

belong to several categories, we should compute the average rating finally. The basic steps of rating predict algorithm are as follows:

Inputs: User-Item rating matrix  $R$ , Item Category

Outputs: The predicted ratings for no rating items

Step1: According to the Item Category  $C$ , the User-Item rating matrix  $R$  can be divided into  $k$  part, that is:  $R_{|U| \times |I|} = R_1 \cup R_2 \cup \dots \cup R_k$ .  $R_j$  represents the  $j$ -th rating matrix.

Step2: On the basis of the above, using the Cosine-based similarity to compute the similarity between items  $i$  and  $j$  like this formula.

$$\text{sim}(i, j) = \cos(i, j) = \frac{\sum_{c=1}^{|U|} R_{c,i} \cdot R_{c,j}}{\sqrt{\sum_{c=1}^{|U|} R_{c,i}^2} \sqrt{\sum_{c=1}^{|U|} R_{c,j}^2}} \quad (1)$$

where  $R_{c,i}$ ,  $R_{c,j}$  represents the rating of the user  $c$  for the item  $i$  and  $j$ , respectively.

Step3: Assuming that  $j$  is an item has no rating of user  $u$ , in order to predict the rating of its, we should generate a neighbor items set  $M_j = \{I_1, I_2, \dots, I_v\}$ ,  $j \notin M_j$  from Step 2 by descending order.

Step4: Using the items in the nearest neighbor set  $M_j$  and the rating values in user-item rating matrix  $R$ , to predict the ratings of the user  $u$  for item  $j$ , denoted by  $P_{u,j}$  is give by

$$P_{u,j} = \frac{\sum_{i \in M_j} \text{sim}(i, j) \times R_{u,i}}{\sum_{i \in M_j} |\text{sim}(i, j)|} \quad (2)$$

where  $R_{u,i}$  presents the rating of the user  $u$  for the item  $i$  in  $M_j$ .

Step5: If item  $j$  belong to multiple categories, repeat Step 2 to Step 4, then taking the average ratings into  $P_{u,j}$  and save it to the rating matrix  $R$ .

### B. User-Item Rating Matrix Expanded by Prediction Rating

Based on the discussed in above, we can build a new rating matrix  $R'_{|U| \times |I|} = (r_{u,j})_{|U| \times |I|}$  by using the prediction ratings like that.

$$r_{u,j} = \begin{cases} R_{u,j} & \text{if User } u \text{ Rated Item } j \\ P_{u,j} & \text{if User } u \text{ Not Rated Item } j \end{cases}$$

### C. Build User-Tag Correlation Matrix

Before going into further detail, the another notation and definitions required for understanding our approach are introduced as follows.

#### 1) Positive and Negative Items

In general, ratings of a user for items can reflect the user's interest more accurately. The rating scale is fixed as numerical values (e.g., a scale of 1-5). Since each user would have his/her own rating behavior, we can classify the items into two parts: a set of positive items and a set of negative items [13].

In this paper, we use the ratings of  $R'$  to form the set of positive and negative items for user  $u$ , which are defined as  $Pos(u)$  and  $Neg(u)$ , respectively such that :

$$Pos(u) = \{i \in I \mid r_{u,i} \geq \bar{r}_u\},$$

$$Neg(u) = \{i \in I \mid r_{u,i} < \bar{r}_u\},$$

in which  $\bar{r}_u$  represents the average rating of user  $u$ .

#### 2) Calculating weights of tags

In our study, we associate a weight of tags with a user's rating, rather than the well-known  $tf-idf$  weight. The weight of tag  $t$  for user  $u$  can be measured by the related items rating of user  $u$ .

Formally, the weight of tag  $t$  annotated in item  $i$  for user  $u$ , can denoted as  $w_{u,i}(t)$ , is computed by:

$$w_{u,i}(t) = \frac{r_{u,i}}{\sqrt{\sum_{j=1}^{|I|} r_{u,j}^2}} \quad (3)$$

Because a tag may appear in several items with different weights, we compute the mean weight of the tag in the set of positive items and negative items, respectively:

$$\omega_{u,t}^{pos} = \frac{1}{|I_u^{pos}(t)|} \times \sum_{j \in I_u^{pos}(t)} w_{u,j}(t) \quad (4)$$

$$\omega_{u,t}^{neg} = \frac{1}{|I_u^{neg}(t)|} \times \sum_{j \in I_u^{neg}(t)} w_{u,j}(t) \quad (5)$$

where  $I_u^{pos}(t)$  and  $I_u^{neg}(t)$  are respectively the set of positive and negative items rated by user  $u$  containing tag  $t$ . Finally, the weight of tag  $t$  for user  $u$ , denoted as  $\omega_{u,t}$ , can be illustrated by the following formula .

$$\omega_{u,t} = \begin{cases} (\omega_{u,t}^{pos} + \omega_{u,t}^{neg}) / 2, & \text{if } t \in I_u^{pos}(t), t \in I_u^{neg}(t) \\ \omega_{u,t}^{pos}, & \text{if } t \in I_u^{pos}(t), t \notin I_u^{neg}(t) \\ \omega_{u,t}^{neg}, & \text{if } t \in I_u^{neg}(t), t \notin I_u^{pos}(t) \end{cases} \quad (6)$$

### 3) User-Tag Correlation Matrix

Based on the mentioned in above, let  $UT_{|U| \times |T|} = (\omega_{u,t})_{|U| \times |T|}$  be a user-tag correlation matrix, in which  $\omega_{u,t}$  represents the weight of tag  $t$  for user  $u$ , that can be computed in the equation (6).

### D. Neighborhood Forming

Neighborhood forming is to generate a set of like-minded peers for a target user  $u_i \in U$  or a set of similar peer items for an item  $p_i \in P$  [12]. In our study, we identify the best neighbors based on the weights for tags, that is why we build the user-tag correlation matrix  $UT$ . Differing from the previous work, rating information is embedded into the tags

when we compute the similarities between users, rather than frequency based weights for the tags.

In order to find  $K$  similar neighbors, various kinds of proximity computing approaches such as cosine similarity and Pearson correlation can be used [12]. Cosine similarity is popularly used to calculate the similarity of two vectors, it also be used in our approaches. Since the vector of tags with weights in  $UT$  is used to represent each item and the preferences of each user, the similarity between users can be measured through calculating the similarity of their weighted tag vectors.

In our method, the similarity between two users  $u_i$  and  $u_j$  is measured by the cosine similarity, that is defined as:

$$sim(u_i, u_j) = \frac{\sum_{t \in T_{u_i, u_j}} \omega_{u_i, t} \times \omega_{u_j, t}}{\sqrt{\sum_{t \in T_{u_i, u_j}} \omega_{u_i, t}^2} \sqrt{\sum_{t \in T_{u_i, u_j}} \omega_{u_j, t}^2}} \quad (7)$$

where  $T_{u_i, u_j}$  refer to the set of tags both in relevant to user  $u_i$  and  $u_j$ .  $\omega_{u_i, t}$  and  $\omega_{u_j, t}$  are respectively the weights of tag  $t$  for user  $u_i$  and  $u_j$ .

Using the similarity measure approach, we can generate the neighborhood of the target user  $u$  by sorting the similarity value in descending order. Formally, the neighborhood of user  $u_i$ , is denoted as:

$$Neigh(u_i) = \{u_j \mid u_j \in \max K\{sim(u_i, u_j)\}\}$$

where  $\max K\{\}$  is used to get the top  $K$  values.

#### E. Recommendation Generation

After generating the set of neighbors, we can learn the historical behavior of neighbor users to predict the target user's fond items. Generally speaking, a set of items that are most frequently

rated or tagged by the neighbors of the target user or the most similar to the target user's rated items will be recommended to the target user [12].

#### 1) The Generation of Candidate Item Set

We assume that the target user  $u$  prefer the items that his neighbors  $Neigh(u)$  prefer, so we generate the candidate item set as follows.

- For each user  $c \in Neigh(u)$ , find the rated items in set  $Pos(c)$ , which contains the fond items of neighbor user  $c$ .
- Delete the items that user  $u$  has rated from  $Pos(c)$ , form the candidate item set, denoted as  $item(u, c, I)$ .
- Combine all the candidate item set of user  $u$ , denoted as  $Can(u, I) = \cup_{c \in Neigh(u)} item(u, c, I)$ .

#### 2) Calculating the Similarity between Items

Define  $I_{u, T}$  is the set of items that  $u$  rated in the past  $T$ , which can reflect the user's current interest to some extent. To ensure the number of items in  $I_{u, T}$  not too small, we can adjust the  $T$  like that. If the number of items in  $I_{u, T}$  is less than 10, we can let  $T = 2 * T$ . Finally, return the set  $I_{u, T}$  and  $T$ .

So, computing the item similarity between the set  $I_{u, T}$  and  $Can(u, T)$  can find the items that the target user  $u$  may prefer. Since an item  $i$  may similar to several items, and has different similarity value. Thus, we can compute the average similarity value as the item's weights for the target user. Formally, it can be computed by :

$$w(u, i) = \frac{sim(i, I_{u, T})}{|I_{u, T}|} = \frac{\sum_{i \in Can(u, T), j \in I_{u, T}} sim(i, j)}{|I_{u, T}|} \quad (8)$$

where  $sim(i, j)$  represents the similarity between item  $i$  and  $j$ . Differing from the similarity calculation in formula (1), the item  $i$  and  $j$  may belong different category, we can use Jaccard formulato compute it as follows:

$$sim(i, j) = \frac{|T(i) \cap T(j)|}{|T(i) \cup T(j)|} \quad (9)$$

where  $T(i)$  and  $T(j)$  represent the number of tags that related to the item  $i$  and  $j$ , respectively.

The top-N items with high similarity scores in formula (8) will be recommended to the target user.

## IV. EXPERIMENT DESIGN

### A. Description of the Data Set

The MovieLens dataset is used in this experiment, which is the most popular dataset used by many scholars to do the collaborative filtering research. It is publicly provided by the GroupLens site (<http://grouplens.org/datasets/movielens/>).

In our experiment, we select the MovieLens 10M as our dataset, it contains 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users. The rating scale ranges from 1 to 5 in which higher ratings indicate greater preference. All movies belong to 19 classes.

### B. Experiments Setup

To evaluate the proposed approaches, we divided the dataset into a training set and a test set randomly. In the dataset, 80% data were randomly used as the training set while rest 20% data were selected as the test set. To ensure the accuracy of experimental results and eliminate the impact of accidental factors, the experiments were repeated five times with different the training/test set. Finally, the result values of this experiment are the averages of the five runs results.

### C. Evaluation Metrics

In this paper, the prediction accuracy of top-N recommendations is evaluated by precision and hit-rank.

- *Precision*: It is used to assess the ratio of the recommended list of items that were also contained in the test set. It is defined as follows:

$$precision = \frac{\sum_{u \in U} |TopN(u) \cap Test(u)|}{\sum_{u \in U} |TopN(u)|}$$

where  $TopN(u)$  is the set of top-N items that recommended to user  $u$ , and the  $Test(u)$  is the set of items rated by user  $u$  in the test data.

To computing the overall precision for all users in the test set, we compute it by averaging the personal precision of each user.

- *Average Reciprocal Hit-Rank (ARHR)*: It was introduced by Deshpande and Karypis (2004) [3], which can be used to assess the hit item's position in the recommended list. Generally speaking, a hit that occurs in the first position is better than a hit that occurs in the N-th position. So, we can give the evaluation according to the hits positions. If  $h$  is the number of hits that occurred at positions  $p_1, p_2, \dots, p_h$  with in the top-N lists (i.e.,  $1 \leq p_i \leq N$ ), then the average reciprocal hit-rank is defined as :

$$ARHR = \frac{1}{n} \sum_{i=1}^h \frac{1}{p_i}$$

That is, hits that occur earlier in the top-N lists are weighted higher than hits that occur later in the list [3].

### D. Results and Discussions

In our proposed approaches, we have the parameters  $T$ . To test the value of  $T$  how to impact on the quality of recommendation, we set the range of parameters  $T$  from 5 to 30 days, and the number of recommend items  $N$  is 10.

As can be seen from Fig. 1, it illustrated that the time window  $T$  between 10 to 15 days has the higher precision, meaning that the recommended effect is good. Also, we can see the precision is decreased when the value of  $T$  is larger than 15, which reflect the large time window will not be able to catch the user's current interest, thus a low precision may be caused.

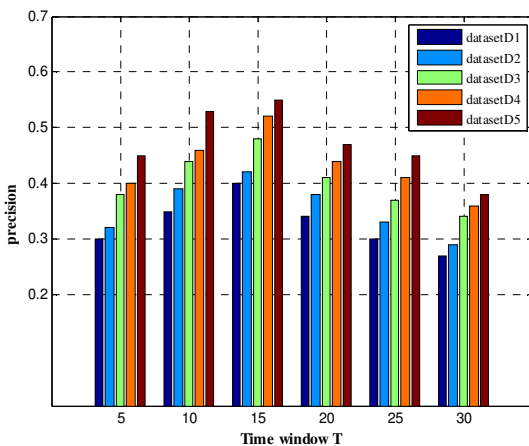


Fig. 1 The precision change with  $T$  ( $N = 10$ ) for each testing dataset

To verify the effectiveness of the algorithm proposed in this paper, we compare the performance of the algorithm with the User-based collaborative filtering algorithm (UBCF), Item-based collaborative filtering algorithm (IBCF) and Tag-based collaborative filtering algorithm (TCF). The experimental results are shown in Fig. 2 and Fig. 3.

Fig. 2 shows us that the relationship between the number of recommend items and the recommend precision. Obviously, different algorithms have the different performances, but the overall trend is

same. With the recommended number  $N$  is rising, the precision of the algorithm is decreasing slowly.

The recommended number of item for 5 ~ 10 has the higher recommendation precision, which means the more recommend items be hit in the test set. Compared with the TCF, IBCF, UBCF algorithm, our algorithm has the highest precision, and the drop speed is also slow. Therefore, the proposed algorithm TARBCF in this paper can capture the user's interests more sensitive, and has more higher credibility.

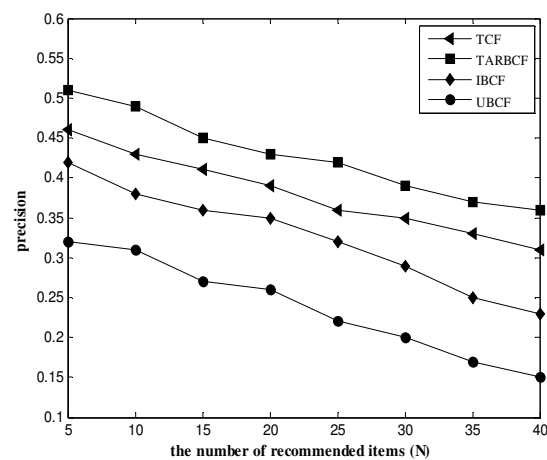


Fig. 1 The average precision comparison of the algorithms

Fig. 3 shows us the performance of different algorithms in the ARHR. From the curve in Fig. 3,

we can see that with the increase of number of recommended resources  $N$ , the ARHR of each algorithm are improved, and the proposed algorithm has the highest ARHR than other schemes. This fully shows that the items list recommended by our algorithm is more fit the needs of the target user.



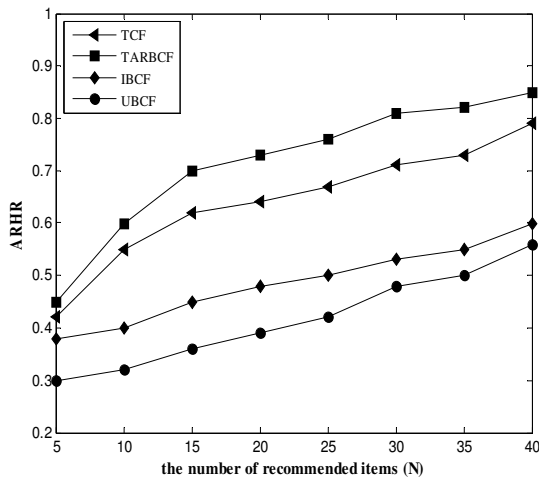


Fig. 3The average ARHR comparison of the algorithms

In a word, the proposed algorithm in this paper has a better performance than others both in the precision and ARHR evaluation metrics. It not only can find the users love items, but also can grasp the degree of user's love. The quality and effect of its recommendations were superior to other solutions.

## V. CONCLUSIONS

In this paper, we presented an improved CF algorithm. To weak the effect of data sparsity in the dataset and make the best use of ratings, a rating predict algorithm based on item category is introduced to predict the ratings of the no rated items, which can reflect the user's hidden interest to some extent. To improve the quality of recommendations, we further build a user-tag correlation matrix by incorporating with ratings and tags, and use it to generate neighborhoods of the target user. To make the recommend items meet the target user's current interest, we also introduce the time window into the recommendation generation phase. To evaluate the proposed algorithm, an experiments based on MovieLens dataset has been conducted. The experiment results show that the proposed algorithm outperforms other traditional algorithm, and has a better quality of recommendation.

In the future, the timestamp of records can be taken into account to track the change of user interests, which can make the recommendation algorithm obtain a higher performance.

## ACKNOWLEDGMENT

I would like to express my gratitude to all those who helped me to complete this research, especially Professor Wang Huijin, my supervisor, for his constant encouragement and guidance. I would also like to thank other members in our Computer Science Lab for their supports.

## REFERENCES

- [1] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 2005:734--749.
- [2] Zhang, et al. "An Improved Collaborative Filtering Algorithm Based on User Interest." *Journal of Software* 9.4(2014).
- [3] Deshpande, M. "Karypis G: Item-Based Top-N Recommendation Algorithms." *Acm Transactions on Information Systems* 22.1(2004):143--177.
- [4] Hu, Jinming. "Application and research of collaborative filtering in e-commerce recommendation system." *Computer Science and Information Technology (ICCSIT)*, 2010 3rd IEEE International Conference on IEEE, 2010:686-689.
- [5] Su, Xiaoyuan, and T. M. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques." *Advances in Artificial Intelligence* 2009.2009(2009).

- [6] Deng AL, Zhu YY, Shi BL. A collaborative filtering recommendation algorithm based on item rating prediction. Journal of Software, 2003,14(9):1621~1628. <http://www.jos.org.cn/1000-9825/14/1621.htm>.
- [7] Ba, Qilong, X. Li, and Z. Bai. "Clustering collaborative filtering recommendation system based on SVD algorithm." Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on IEEE, 2013:963-967.
- [8] Chen, Yongping, S. Yang, and Y. Liu. "Time-weighted collaborative filtering algorithm based on item content and rating." Journal of Suzhou University of Science & Technology(2013).
- [9] Ching-man Au Yeung ., Nicholas Gibbins ., and N. Shadbolt. "Multiple Interests of Users in Collaborative Tagging Systems." Weaving Services & People on the World Wide Web(2009):115 - 118.
- [10] Nakamoto, Reyn Y., et al. "Reasonable tag-based collaborative filtering for social tagging systems." Proceedings of the 2nd ACM workshop on Information credibility on the web ACM, 2008.
- [11] Wang, J., Clements, M., Yang, J., Vries, A. P. D., & Reinders, M. J. T. (2010). Personalization of tagging systems. Information Processing & Management An International Journal, 46(1), 58-70.
- [12] Liang, Huizhi, et al. "Connecting users and items with weighted tags for personalized item recommendations.." Proc of Ht'(2010):51-60.
- [13] Kim, Heung Nam, et al. "Collaborative user modeling with user-generated tags for social recommender systems." Expert Systems with Applications 38.7(2011):8488–8496.

# Performance of data mining algorithms in unauthorized intrusion detection systems in computer networks

Hadi Ghadimkhani, Ali Habiboghli\*, Rouhollah Mostafaei

Department of Computer Science and Engineering, Islamic Azad University, khoy Branch, Khoy, Iran

Corresponding Author : Ali Habiboghli

\*\*\*\*\*

## Abstract:

In this article, we first normalized data by using random sampling conducted and chose a limited number of data. Then we determined the two types of clusters using hours and non-working hours and hours, and we implement the clusters ID3 and Decision tree algorithms. The results of our proposed method show that ID3 algorithm for high power at sorting and classification acts better than Decision tree. The purpose of this paper is to compare the performance of these two algorithms.

**Keywords —Intrusion Detection, Network, Data Mining.**

\*\*\*\*\*

## I. INTRODUCTION

In today's world, computers and computer networks connected to the Internet will play a major role in communication and information transfer. In the meantime, jobber with access to important data or the information centres of other people and with the intent to influence or pressure, or even to disassemble order systems, have adopted the act to the computer systems. Thus the need to protect the information security and maintain the effectiveness in computer networks that communicate with the outside world is necessary [7].

An Intrusion Detection System (IDS) is a security technique attempting to detect various attacks. It has been identified mainly two techniques, namely misuse detection and anomaly detection [1, 2, 8, 9].

With the growth of information technology, network security is proposed as an important issue and very large challenge [7]. Intrusion detection system is the main component of a secure network. Traditional intrusion detection systems cannot adapt with new attacks, therefore Data mining-based intrusion detection systems offered today. Here we use data mining methods and reduced features and Clustering and classification techniques on our database.

Misuse detection, also called signature-based detection, attempts to model abnormal behaviour and normally focuses on the known attacks. It uses a descriptive language to delineate the characteristics of the known attacks and to construct the corresponding attack signatures. However, it may not be able to alert the system administrator in case of a new attack. Anomaly detection attempts to model normal behaviour. Any events which deviates the normal usage patterns are considered to be suspicious. It constructs the profile of user behaviour or status of network traffic and compares the observed behaviour with the stored profile to determine whether an attack action occurs. The anomaly detection approach may have the advantage of detecting previously unknown attacks over the misuse detection approach. However, it may suffer from false alarm problem and radically changed user behaviours [3].

Except the above taxonomy of intrusion detection techniques, IDS is classified into host-based and network-based IDS by their defensive scopes [10].

## II. DESCRIPTION OF PROBLEM

### A. Network intrusion detection system

A network intrusion detection system or Network-Based IDS, Briefly to say NIDS, In fact, is a variety of network intrusion detection system [4] which is

connected to the network and in this way, monitoring network traffic and provides reports. Methods of making such systems are usually behind or in front of the firewall network. In Figure 1 you can see how to position the network intrusion detection system. Always the best location for a network intrusion detection system NIDS, is outside the network firewall.

Jing and Papavassiliou [5] propose a new network traffic prediction methodology based on the frequency domain traffic analysis filtering.

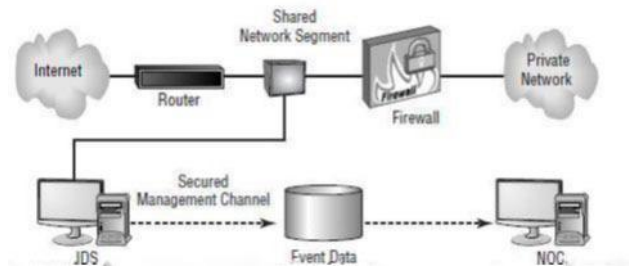


Figure 1. Wrapping method NIDS in network shows how to analyse network traffic.

## B. Architectural types of intrusion detection systems

Various architectural intrusion detection systems include:

- Host-based intrusion detection system (HIDS)
- Network-based intrusion detection system (NIDS)
- Distributed Intrusion Detection System (DIDS) [6].

### B.1. Host-based intrusion detection system (HIDS)

The system is responsible to detect unauthorized activity on the host computer. Host-based intrusion detection system can detect attacks and threats to critical systems (Includes access to files, Trojans and etc.) that are not detectable by the network-based intrusion detection systems. HIDS due to their location on the host to be monitored are notified all types of additional local information by

implement security (Including system calls, change system files and system connections). This when combined with the communications network, may provide a good data to search for possible event [6].

### B.2. Network-based intrusion detection system (NIDS)

NIDS name is derived from the fact that of where it is located, the entire network is monitored.

Network-based intrusion detection systems adapt to detect unauthorized intrusions before they reach critical systems. NIDS are mostly formed two monitors (sensors) and factor. These are often installed behind the firewall and other access points to detect any unauthorized activity [6].

### B.3. Distributed Intrusion Detection System (DIDS)

These systems are composed of several HIDS or NIDS or the combination of these two with a central management station.

The statements that any IDS are available on the network send their reports to the central management station. Central Station is responsible to examining the reports and alert security officer. This central station also is responsible to updates detection rule base of each IDS on the network [6].

## III. PROPOSED METHOD

For network analysis, a method used that to the help the centrality of the concept of dependency, directed graph become hierarchical graph. Using hierarchical structure can be easily distinguished between leaders and followers till the terrorist network to be built. New ideas for measuring the centrality dependence is much to do hierarchical model, for this idea shows nodes that are entirely dependent on a particular node. How to calculate the central of attachment is in Formula 1 [11, 12].

The central of attachment

$$DC_m = \sum_{m \neq p, p \in G} \frac{d_{mn}}{N_p} + \Omega \quad (1)$$

That

$N_p$  = Number of nodes per cluster

And

$m, n$  = The distance between two nodes

Using measurements of network performance is well represented the effect of each node of the graph. The removal of a node, the further decrease network performance ,the importance of their presence is tied more. The calculation of network performance is displayed in formula 2.

Formula 2, the  $d_{ij}$  value represents the shortest distance between two nodes of  $j$  and  $i$  [11, 12].

Efficiency centrality

$$E(G) = \frac{\sum_{i \neq j} e_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} \quad (2)$$

That

$i, j$  = The distance between the node

And

$N$  = Number of nodes

And

$d_{i,j}$  = The shortest distance between two nodes

Central Profile of organizational role specifies the role of the network. Corporate roles Profile will be counted by using centralized network performance. If the central organizational role profile, plotted on the coordinates, the top of the axis represents leaders and below the horizontal axis, represents the X role of soldiers and less important members of the group.

How to calculate the profile of organizational role are shown in Formula 3 [11,12].

Organizational designation index

$$PRI = E(G) - E(G - node_i), i = 1, 2, \dots, N \quad (3)$$

That

$E(G)$  = The centrality of the total performance

And

$node_i$  = Central node performance

## IV. EXPERIMENTAL RESULTS

### A. Selection step and data collection

Here we use a limited number of data which are includes features as date, time, user, pc, activity.

### B. Data preparation

At this stage, first discrete data is carried out.

Activity(count)	activity	date	row
115	logon	2010/04/01	1
28	logoff	2010/04/01	2
34	logon	2010/05/01	3
34	logoff	2010/04/01	4

Table 1- the number of log off, log on by date

### C. Sampling

Here, first we have Categories the data based on date and connect and disconnect rate:

So, as you can observe according to table1 and figure 2 (2010/04/01), numbers of logon are more than logoff and it can be cited that this date can be mystification for us.

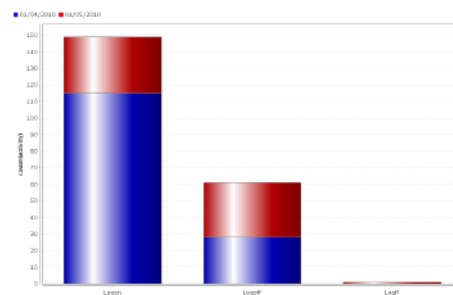
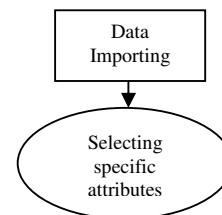


Figure 2. Frequency of logon and logoff according time

In Figure 3 you can see the ID3 algorithm implementation process.





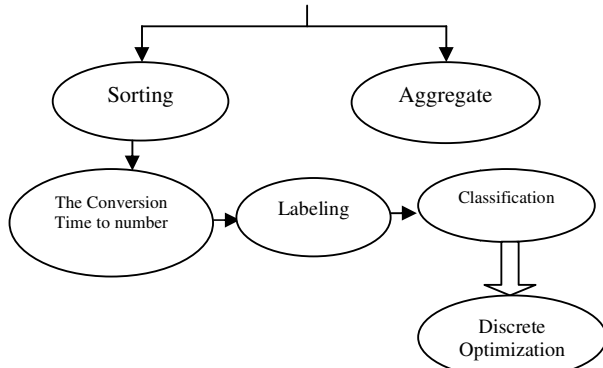


Figure 3. Implementation Process of ID3 Algorithm

Decision tree algorithm implementation process can be seen in Figure 4.

File CSV is the study input that was taken from a site that was said and enters into Rapid Miner software for analysis.

According to table 2, we have turned the clocks into numbers to be able to do calculations more comfortably. For this aim, first we have sorted the date and time ascending.

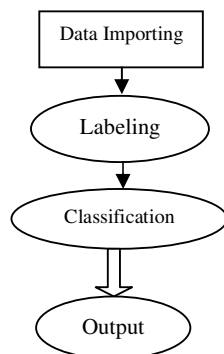


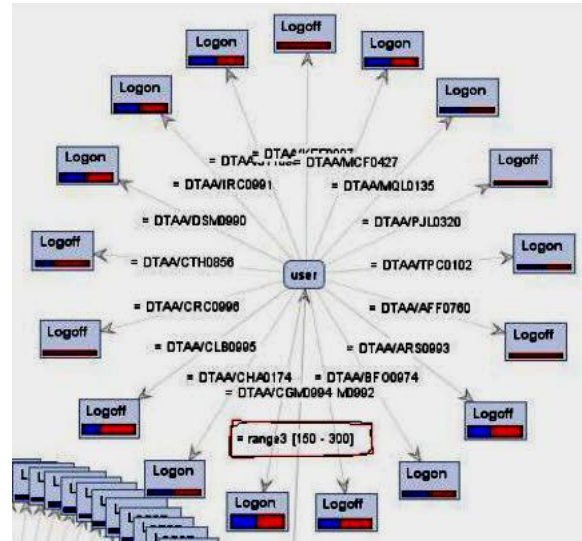
Figure 4. Implementation Process of Decision Tree Algorithm

Numerical Conversion	Time	date	row
0	07:00:10	2010/04/01	1
1	07:00:53	2010/04/01	2
2	07:01:09	2010/04/01	3
3	07:01:52	2010/04/01	4
4	07:02:27	2010/04/01	5
5	07:03:13	2010/04/01	6

6	07:03:17	2010/04/01	7
7	07:03:23	2010/04/01	8
8	07:03:52	2010/04/01	9
9	07:04:05	2010/04/01	10

Table 2-Conversion time to number and sort them ascending

Figure 5-Clustering Chart id3



Thus, as you can infer according to figure 5, first hours of day, all users are logon, but in range [150-300] that is related to last hours of work, it can be cited that, all of users has gotten logon same as logoff hours.

But, you can understand with care that DTAA/CRC0996 - DTAA/KEE0997- DTAA/AFF0760 users are logoff perfectly in these work hours, for more accurate security, manager can pay attention to clustering algorithm written part that as sample has been displayed in figure 6.

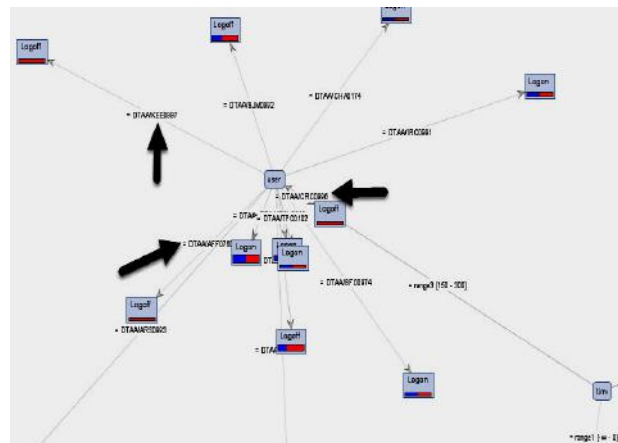


Figure 6. graph clustering of ID3

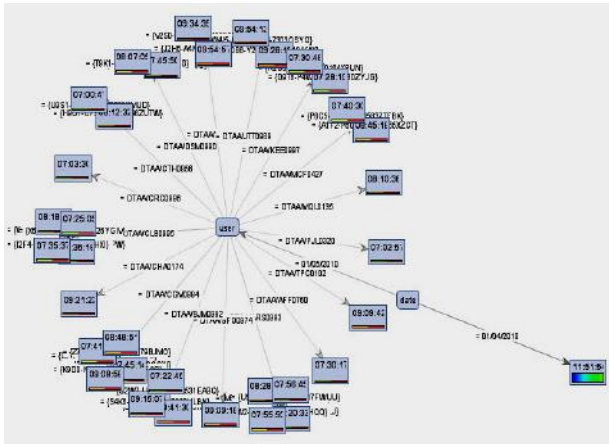


Figure 7- Decision tree diagram

As you can see in Figure 7, in this algorithm can be seen different id to a particular user on 01/05/2010 but this algorithm has not done well classification to date 01.04.2010 that could be a weak point of the algorithms. If you stand on any of the classes built, you can see the members of the class and through this comparison, we can say that the classification ID3 is much better than decision tree.

## V. CONCLUSION

To establish complete security in a computer system, in addition to firewalls and other intrusion prevention systems called intrusion detection systems (IDS) are needed to see if the attacker of firewall, antivirus, other security was passed and signed, it detects and think of ways to deal with it. In this study, we used data information network to possible data mining and intrusion detection or any attack or any unacceptable action which was conducted according to regulations and laws. We first normalized data analysed using random sampling and volume data on 300 others. Then ID3 and decision tree algorithm on them was implemented. ID3 algorithm acts better than decision tree in packing and sorting, and this is one of the strengths of ID3 algorithm, however, Decision Tree cannot sorted the information, but in processing acts better than ID3. Decision tree

algorithm has high error probability in the case of low and high category, Because of the limited number of data used in this article.

## REFERENCES

- [1].C. M. Chen, Y. L. Chen and H. C. Lin, "An efficient network intrusion detection", Computer Communications (33), pp: 411–424, 2111.
- [2].T. Verwoerd, R. Hunt, Intrusion detection techniques and approaches, Computer Communications 25 (15) (2002) 1356–1365.
- [3].E.Lundin and E.Jonsson, "Anomaly-based intrusion detection: privacy concerns and Otherproblems", Journal of Computer Networks, volume (3), number (4), pp: 623–640, 2111.
- [4].P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava and P. N. Tan, "Data Mining For Network Intrusion Detection", Computer Science Department, 211 Union.Street SE,4-102, EE/CSC Building University of Minnesota, Minneapolis, MN 11411, USA, 2012.
- [5].J. Jing, S. Papavassiliou, Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies, Computer Communications 29 (10) (2006) 1627–1638.
- [6].EhsanMalekian, An attacker on the network and coping strategies, Nas Publisher, 2009, in Persian
- [7].Masoud Sotoudeh Far, Network intrusion detection based on adaptive fuzzy rules, master thesis, 2004, in Persian.
- [8]. E. Biermann, E. Cloete, L.M. Venter, A comparison of intrusion detection systems, Computer and Security 20 (8) (2001) 676–683.
- [9]. J.M. Estevez-Tapiador, P. Garcia-Teodoro, J.E. Diaz-Verdejo, Anomaly detection methods in wired networks: a survey and taxonomy, Computer Communications 27 (16) (2004) 1569–1584.
- [10]. H.S. Venter, J.H.P. Eloff, A taxonomy for information security

technologies, Computer and Security 22 (4)  
(2003) 299–307.

- [11]. Memon Nasrullah and Henrik Legind Larsen, Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks. In the proceedings of ARES 2006: The First International Conference on Availability, Reliability and Security, Vienna, Austria, IEEE Computer Society, pp. 906-913, 2006.
- [12]. Memon Nasrullah and Henrik Legind Larsen, Practical Algorithms of Destabilizing Terrorist Networks. In the proceedings of IEEE Intelligence Security Conference, San Diego, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3976: pp. 398-411, 2006.

# Vibration Analysis of a Horizontal Washing Machine, Part IV: Optimal Damped Vibration Absorber

**Galal Ali Hassaan**

Department of Mechanical Design & Production, Faculty of Engineering,  
Cairo University, Egypt

\*\*\*\*\*

## Abstract:

This is the 4<sup>th</sup> research paper investigating the vibration of a horizontal washing machine. In this paper a dynamic vibration absorber is attached to the machine drum. The dynamic system is modeled and the dynamic absorber is assumed to have a known mass and damping coefficient. Only the absorber stiffness is tuned using MATLAB optimization toolbox. The drum vibration velocity is used as an objective function to attain a value compatible with ISO standard 10816. The isolation efficiency is used as a functional constraint to succeed in isolating the large unbalanced rotating force of the laundry during the spinning cycle of the washing machine. The simulation results are amazing. The proposed approach can reduce the vibration velocity to less than 0.7 mm/s RMS and increase the isolation efficiency to greater than 99.7 %.

**Keywords —** Horizontal washing machine, vibration control, dynamic vibration absorber, absorber tuning.

\*\*\*\*\*

## I. INTRODUCTION

Vibration of washing machines during the spinning cycle is something undesirable since it has high levels usually beyond the levels recommended by international standards. The author investigated the vibration of horizontal washing machines in a series of research papers. In the present paper the author applies a dynamic vibration absorber to limit the vibration velocity of the drum and increase the isolation efficiency of the machine suspension to higher levels to decrease noise generated during the spinning cycle.

Miller (2003) investigated the nonlinear mechanical properties of absorbers for potential application in adaptive-passive tuned vibration absorbers. He used an adaptive-passive tuned vibration absorber capable of varying the absorber resonant frequency in the range 45 to 211 Hz [1]. Liu and Liu (2005) revisited a classical problem about optimum damped vibration absorber. They have applied Brock's approach to different type of damped vibration absorber named model B (when

the main system is damped and the absorber damper is connected to the ground) and found its optimal parameters [2]. Thompson (2007) considered a mass-spring absorber system to attenuate structural waves in beams. He investigated the parameters controlling the behaviour of the absorber and developed a formula allowing optimization of its performance [3].

Krenk and Hogsberg (2008) presented a design procedure for tuned mass absorbers mounted on structures with structural damping. They found an accurate explicit approximation for the optimal damping parameter of the absorber and the resulting damping ratio of the response [4]. Najafi, Ashory and Jamshidi (2009) suggested different models for the vibration absorbers to assign the optimal one for SDOF vibration suppression. They used genetic algorithm to optimize the best absorber model [5]. Lin and Coppola (2010) studied the optimal design of the damped dynamic vibration absorber for damped primary systems. They applied two numerical approaches by solving a set of nonlinear equations by the Chebyshev equioscillating theorem

and by minimizing a compound objective function subject to a set of constraints [6]. Brown and Singh (2011) formulated a minimax problem to determine the parameters of vibration absorber by minimizing the maximum motion of the primary mass over the exciting frequency domain. They could minimize the main mass displacement magnitude to a value lower than methods available in the literature [7].

Fang, Wang and Wang (2012) investigated a minimax design of damped dynamic vibration absorber for a damped primary system to minimize the vibration amplitude. They illustrated the advantage of their proposed method through numerical simulations [8]. Huang and Lin (2014) designed a dynamic vibration absorber called periodic vibration absorber for mechanical systems subjected to periodic excitation. Their results showed that the periodic vibration absorber could be a very effective device for vibration reduction of mechanical systems subjected to periodic excitation [9].

Kamran, Rezazadeh and Ghaffari (2015) investigated reducing the unwanted vibration in machine tools. They proposed a algorithm to achieve the optimal parameters of the vibration absorber. They evaluated the effectiveness of the proposed algorithm and the designed vibration absorber through comparing the vibration amplitude of the machine tool in the presence and absence of the absorber [10]. Abdelhafiz and Hassaan (2015) investigated using an adaptive tuned vibration absorber to maximize the vibration attenuation of the main vibrating system. Their tuning condition was used to track the exciting frequency. They showed that the frequency response of the main system could be reduced by 10 % [11].

## II. THE WASHING MACHINE PHYSICAL MODEL

A physical model of the horizontal washing machine is shown in Fig.1 [12], [13]:

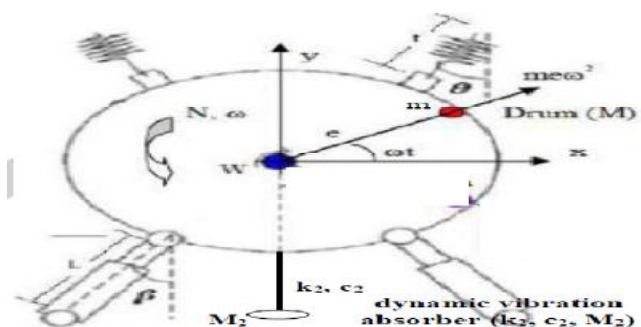


Fig.1 Physical model of the horizontal washing machine [12], [13].

The drum mass is  $M$ , the laundry mass is  $m$  and its eccentricity is  $e$ . The suspension stiffness is  $k$ , damping coefficient is  $c$ , spring inclination is  $\theta$  and damper inclination is  $\beta$  with the vertical axis. The rotor speed is  $N$  (rev/min) and  $\omega$  (rad/s).

A dynamic vibration absorber consisting of a cantilever of stiffness  $k_2$ , damping coefficient  $c_2$  and a lumped mass of mass  $M_2$  is attached to the drum at its bottom as shown in Fig.1.

## III. VIBRATING SYSTEM MATHEMATICAL MODEL

The mathematical model of the system is constructed as follows:

1. The washing machine drum is the main vibrating system of motion  $x$  (assumed as a SDOF system [14]). It has a mass  $M$  and its suspension of parameters  $k$  and  $c$ .
2. The absorber of lumped mass  $M_2$  and stiffness  $k_2$  is attached to the drum. Its mass vibrates with a dynamic motion  $x_2$ .
3. The system is a two degree of freedom one excited by the inertia force of the laundry ( $me\omega^2$ ).
4. It is assumed that the suspension and absorber parameters have linear characteristics.
5. The mathematical model is derived by drawing the free body diagram of each mass and applying Newton's second law of motion. This step yield two differential equations in  $x$  and  $x_2$ .
6. The two differential equations are written in matrix form as follows:

$$M\ddot{x} + C\dot{x} + Kx = F_0 e^{j\omega t} \quad (1)$$

Where:  $M$  = mass matrix =  $\begin{bmatrix} M & 0 \\ 0 & M_2 \end{bmatrix}$

$C$  = damping matrix =  $\begin{bmatrix} c' + c_2 & -c_2 \\ -c_2 & c_2 \end{bmatrix}$

$K$  = stiffness matrix =  $\begin{bmatrix} k' + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix}$



$$\mathbf{X} = \text{peak amplitude vector} = \begin{bmatrix} X \\ X_2 \end{bmatrix}$$

$$\mathbf{F}_0 = \text{exciting force vector amplitude} = \begin{bmatrix} m\omega^2 \\ 0 \end{bmatrix}$$

The main system has an equivalent stiffness,  $k'$  and damping coefficient,  $c'$  given by [12]:

$$k' = 2k(\sin\theta)^2$$

$$\text{and } c' = 2c(\sin\beta)^2$$

The steady state solution of Eq.1 is [15]:

$$\mathbf{x} = \underline{\mathbf{X}} e^{j\omega t} \quad (2)$$

where  $\underline{\mathbf{X}}$  is the amplitude phasor (complex value). Combining Eqs 1 and 2 gives:

$$(\mathbf{K} - \omega^2 \mathbf{M} + j\omega \mathbf{C}) \underline{\mathbf{X}} = \mathbf{F}_0 \quad (3)$$

Eq.3 give the phasor of the vibration amplitudes as:

$$\underline{\mathbf{X}} = (\mathbf{K} - \omega^2 \mathbf{M} + j\omega \mathbf{C})^{-1} \mathbf{F}_0 \quad (4)$$

According to Miller, Eq.4 gives the vibration amplitude of the main mass,  $X$  as [1]:

$$X = m\omega^2 \sqrt{\{N_1 / (D_1 + D_2)\}} \quad (5)$$

Where:

$$N_1 = (k_2 - M_2 \omega^2)^2 + (c_2 \omega)^2$$

$$D_1 = [(k' - M \omega^2)(k_2 - M_2 \omega^2) - k_2 M_2 \omega^2 - c' c_2 \omega^2]^2$$

$$D_2 = \omega^2 \{ [k_1 - (M_1 + M_2) \omega^2 + c' (k_2 - M_2 \omega^2)]^2 \}$$

The vibration velocity of the washing machine has an RMS amplitude given by:

$$V = 0.707 \omega X \quad \text{mm/s} \quad (6)$$

The isolation efficiency of the washing machine isolators,  $\eta$  is [16]:

$$\eta = 100(1 - \text{TR}) \quad (7)$$

Where TR is the transmissibility of the drum vibration given by:

$$\text{TR} = F_{t0} / (m\omega^2) \quad (8)$$

$$\text{Where } F_{t0} = X \sqrt{\{k'^2 + (\omega c')^2\}}$$

#### IV. OPTIMIZING THE VIBRATION ABSORBER

The dynamic vibration is optimized using the following approach:

1. The number of absorber parameters are reduced from three to parameters to only one parameter.
2. The mass of the absorber is set to 2 kg and its structural damping coefficient is set to 10 Ns/m.
3. The left parameter is the absorber stiffness which can be controlled by the cantilever length, cross-sectional area and material.
4. The algorithm used depends of assigning the absorber stiffness  $k_2$  to minimize an objective function subject to a functional constraint and a stiffness boundaries.
5. The MATLAB command '*fmincon*' is used for this purpose [17].
6. The objective function used is the vibration velocity of the machine drum. It is required to minimize the drum vibration velocity.
7. The functional constraint used is the isolation efficiency. It is required to keep the vibration velocity  $\geq 98\%$ .
8. The vibration absorber stiffness is bounded as:  $500 \leq k_2 \leq 5000 \text{ N/m}$ .
9. This optimization procedure is applied for spinning machine N in the range  $400 \leq N \leq 1200 \text{ rev/min}$ .
10. The results of the application of this optimization procedure are given in Table 1 for a machine drum of 30 kg mass and suspension of 5000 N/m isolator stiffness and 130 Ns/m damping coefficient.

TABLE I  
OPTIMIZING THE VIBRATION ABSORBER

Spinning speed (rev/min)	Optimal system performance			
	$k_2$ (N/m)	X (mm)	V (mm/s RMS)	$\eta$ (%)
400	1000	0.0223	0.6605	99.6867
500	1000	0.0160	0.5925	99.8356
600	1000	0.0126	0.5609	99.8979
700	1000	0.0105	0.5434	99.9300
800	1000	0.0090	0.5326	99.9488
900	1000	0.0079	0.5254	99.9608
1000	1000	0.0070	0.5203	99.9690
1100	1000	0.0063	0.5167	99.9748
1200	1000	0.0058	0.5139	99.9791

11. The optimal vibration peak amplitude of the drum as function of the machine spinning speed is shown graphically in Fig.2 compared with that without vibration absorber.

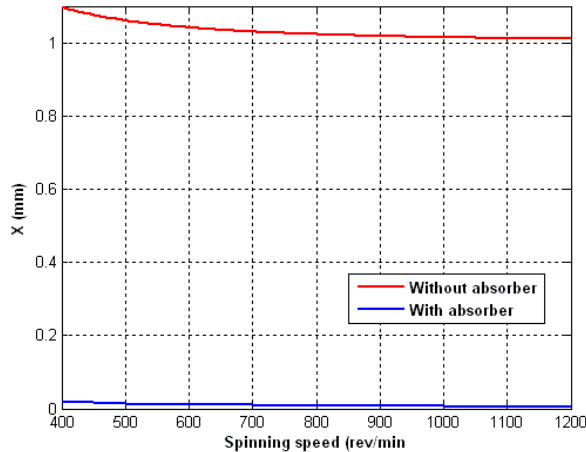


Fig. 1 Optimal vibration amplitude of the washing machine drum.

12. The effect of the spinning speed on the optimal drum vibration velocity in mm/s RMS is shown in Fig.3 for the cases with using a dynamic vibration absorber and without vibration absorber. Good vibration velocity limit is shown according to ISO 10816 [18].

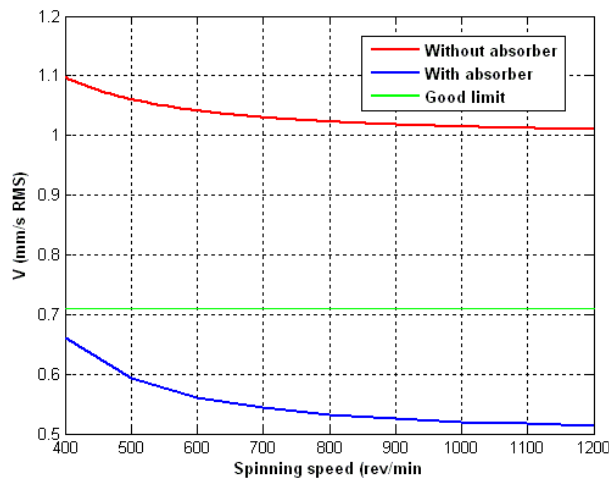


Fig. 3 Optimal vibration velocity of the washing machine drum.

13. The effect of the spinning speed on the optimal isolation efficiency is shown in Fig.4 for the cases with using a dynamic vibration absorber and without vibration

absorber. It shows also the desired minimum isolation efficiency.

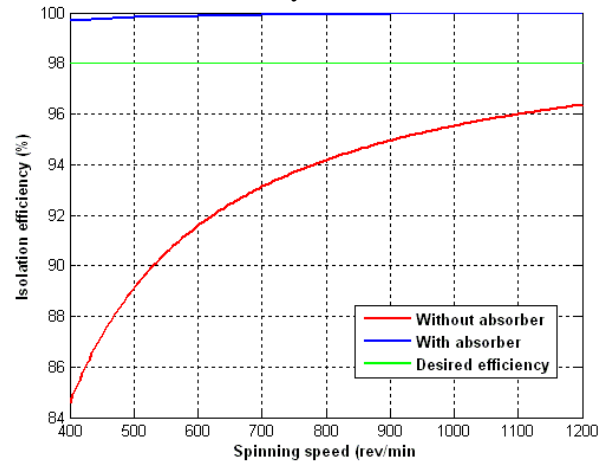


Fig. 4 Optimal isolation efficiency of the washing machine suspension.

## V. CONCLUSIONS

- In order to improve the design of the drum-suspension system of a horizontal washing machine, the parameters are optimized for minimum vibration velocity and maximum isolation efficiency.
- The MATLAB toolbox was used for this purpose.
- Only one design parameter was used which is the isolator stiffness.
- The objective function used was the vibration velocity of the drum in mm/s RMS to be compatible with the ISO requirements.
- The functional constraint used was the isolation efficiency such that it has to be above 98 %.
- A value of 1000 N/m isolator stiffness was found reasonable from point of view of vibration velocity and isolation stiffness.
- The effect of the spinning speed of the washing machine was investigated.
- The optimization approach was very effective since it could reduce the vibration amplitude to less than 0.006 mm at 1200 rev/min spinning speed. It could reduce the drum vibration velocity to less than 0.7 mm/s RMS for spinning speeds  $\geq 400$  rev/min.

- It could increase the isolator efficiency to above 99.68 % for spinning speeds  $\geq 400$  rev/min.

## REFERENCES

- [1] S. Miller, "The development of a nonlinear elastomeric adaptive tuned vibration absorber", M. Sc. Thesis, Faculty of Engineering, Rowan University, USA, October 2003.
- [2] K. Liu and J. Liu, "The damped dynamic vibration absorbers revisited and new result", *Journal of Sound and Vibration*, vol.286, pp.1181-1189, 2005.
- [3] D. Thompson, "The theory of continuous damped vibration absorber to reduce broad-band wave propagation in beams", *ISVR Technical Memorandum No.968*, 52 pages, January 2007.
- [4] S. Krenk and J. Hogsberg, "Tuning mass absorbers on damped structures", *7<sup>th</sup> European Conference on Structural Dynamics*, Southampton, UK, Paper E61, 7-9 July 2008.
- [5] N. Najafi, M. Ashory and E. Jamshidi, "Optimum configuration for vibration absorbers of a SDOF system using genetic algorithm", *Proceedings of the IMAC-XXVII*, Orlando, USA, 10 pages, 9-12 February 2009.
- [6] K. Lin and G. Coppola, "Optimal design of damped dynamic vibration absorber for damped primary systems", *Transactions of the Canadian Society for Mechanical Engineering*, vol.34, issue 1, pp.119-135, 2010.
- [7] B. Brown and T. Singh, "Minimax design of vibration absorbers for linear damped systems", *Journal of Vibration and Sound*, vol.330, issue 11, pp.2437-2448, 2011.
- [8] J. Fang, S. Wang and Q. Wang, "Optimal design of vibration absorber using minimax criterion with simplified constraints", *Acta Mechanica Sinica*, vol.28, issue 3, pp.848-853, 2012.
- [9] S. Huang and K. Lin, "A new design of vibration absorber for periodic excitation", *Shock and Vibration*, vol.2014, Article ID 571421, 11 pages, 2014.
- [10] M. Kamran, G. Rezazadeh and G. Ghaffari, "An investigation on optimal design of dynamic vibration absorber using genetic algorithm", *Science Journal*, vol.36, issue 3, Special Issue, 15 pages, 2015.
- [11] M. Abdelhafiz and G. A. Hassaan, "Tuning condition modification of damped and un-damped adaptive vibration absorber", *International Journal of Computer Techniques*, vol.2, issue 2, pp.170-175, 2015.
- [12] G. A. Hassaan, "Vibration analysis of a horizontal washing machine, PartI: Vibration displacement and velocity", *International Journal of Engineering and Advanced Research Technology*, vol.3, issue 4, 2015, Accepted for publication.
- [13] A. Yorukoglu and E. Altug, "Determining the mass and angular position of the unbalanced load horizontal washing machines", *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, Singapore, pp.118-123, 2009.
- [14] S. Solhar and D. Patel, "Optimization of a drum type washing machine by analytical and computational assessment", *International Journal of Scientific and Engineering Research*, vol.4, issue 6, pp.2759-2763, 2013.
- [15] L. Meirovitch, *Fundamental of Vibrations*, Waveland Press Inc., 2010.
- [16] P. Antonio, "Predicting isolation efficiency", [www.hallani.com](http://www.hallani.com), 13 pages, January 2010.
- [17] P. Venkataraman, *Applied Optimization with MATLAB Programming*, J. Wiley, 2009.
- [18] J. Robichaud, "Reference standards for vibration monitoring and analysis", [www.bretech.com/reference/](http://www.bretech.com/reference/), April 2011.

## BIOGRAPHY



**Galal Ali Hassaan**

- Emeritus Professor of System Dynamics and Automatic Control.
- Has got his B.Sc. and M.Sc. from Cairo University in 1970 and 1974.
- Has got his Ph.D. in 1979 from Bradford University, UK under the supervision of Late Prof. John Parnaby.
- Now with the Faculty of Engineering, Cairo University, EGYPT.
- Research on Automatic Control, Mechanical Vibrations , Mechanism Synthesis and History of Mechanical Engineering.
- Published 10's of research papers in international journals and conferences.
- Author of books on Experimental Systems Control, Experimental Vibrations and Evolution of Mechanical Engineering.
- Chief Justice of the International Journal of Computer Techniques.
- Member of the Editorial Board of some international journals including IJCT.
- Reviewer in some international journals.
- Scholars interested in the authors publications can visit:

<http://scholar.cu.edu.eg/galal>

# Prediction of Heart Disease Using Enhanced Association Rule Based Algorithm

Karandeep Kaur\*, Ms. Poonamdeep Kaur\*\*, Ms. Lovepreet Kaur\*\*\*

\*(Student (Computer Science & Engineering), Guru Nanak Dev Engineering College, Ludhiana)

\*\* (Assistant Professor (Computer Science & Engineering), Guru Nanak Dev Engineering College, Ludhiana)

\*\*\* (Assistant Professor (Information Technology), Guru Nanak Dev Engineering College, Ludhiana)

\*\*\*\*\*

## Abstract:

The risk of coronary illness is increasing at a fast pace and it's been scowling at us for a considerable length of time, making us doubt each subtle element of our confounded way of life decisions, eating regimen and level of physical movement. It's been a main executioner in the West and has now forcefully advanced toward India. As indicated by government information, the pervasiveness of heart disappointment in India because of coronary illness, hypertension, corpulence, diabetes and rheumatic coronary illness ranges from anyplace between 1.3 to 4.6 million, with a yearly frequency of 491,600 to 1.8 million. But because of impreciseness of the diagnosis tools less than 68 per cent of heart diagnosis yield correct results. To make the odds go higher, this research presents a novel algorithm whose key idea is germinated from classical association rule mining.

**Keywords — Association Rule Mining, Heart Disease, UCI Machine Learning.**

\*\*\*\*\*

## I. INTRODUCTION

Mechanical progressions and social insurance mindfulness have driven towards the improvement of tremendous number of human services offices and doctor's facilities. On the other hand, given the high caliber of medicinal services administrations requiring little to no effort is turning into a testing issues inside the developing nations across the globe albeit, numerous nations have made some firm strides towards the guaranteeing that human services administrations are given to everybody. Therapeutic information mining has an incredible capacity for the investigation of the shrouded examples in the current datasets of the medicinal area. Such examples could be used with the end goal of clinical conclusion. A noteworthy test that this Industry's appearances is giving quality administrations at reasonable costs. Quality administration suggests diagnosing an understanding's condition viably, giving fitting medications and observing those medicines all the time. Uncalled for clinical choices may prompt sad results. Alongside diagnosing conditions and giving

suitable medications, healing centres should likewise minimize the expense of clinical tests. A successful method for accomplishing these outcomes is by utilizing proper PC - based data and/or choice emotionally supportive networks. Most doctors' facilities today utilize some kind of data frameworks to deal with their medicinal services or patient information. These frameworks create colossal measures of information as numbers, outlines, writings pictures and so forth. Gigantic measure of information is put away on consistent premise however sadly such information is infrequently used to bolster clinical choice making. There is a substantial number of shrouded data in these information that is to a great extent undiscovered.

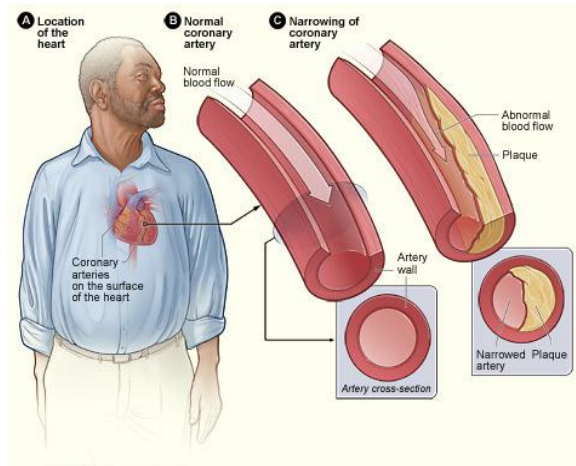


Fig.1: Coronary Heart Disease

Information mining obliges an accumulation of information in a sorted out shape, the information gathered was incorporated in the arrangement of a clinic data framework. The innovation of information mining gives the clients a client arranged methodology towards concealed and novel examples in information. Proficient and powerful robotized coronary illness framework fit for foreseeing coronary illness could turn out to be tremendously useful for the social insurance segment. This research study endeavours to present a point by point investigation of diverse information of Enhanced Association Rule Learning for the expectation of coronary illness using the information comprising of a person's medical history. These mechanizing frameworks will assume a noteworthy part in lessening the general number of tests that a patient needs to take concerning coronary illness.

## II. METHODOLOGY

This paper formulates a novel algorithm based on association rule learning to precisely predict the result of a coronary illness examination and the act of this novel calculation could turn out to be useful for the medicinal specialists and experts for precisely anticipating the coronary illness.

### A. Inputs

13 attributes taken from 270 different observations is the input to the proposed system. The type of each attribute is given below:

Attribute	Type
Age	Quantitative
Sex	Categorical
Chest Pain Type	Categorical
Resting Blood Pressure	Quantitative
Serum Cholesterol	Quantitative
Fasting Blood Sugar	Categorical
Resting Electrocardiographic results	Categorical
Maximum Heart Rate Achieved	Quantitative
Exercise Induced Angina	Categorical
Oldpeak	Quantitative
ST segment	Categorical
Coloured Vessels	Categorical
Thal	Categorical
Result	Categorical

### B. Algorithm

The algorithm of the proposed methodology is given below:

```

Step 1: Scan  $x=1$  to  $n$ 
Step 2: Set  $\text{Min}(\text{Support}) = \mu$ 
Step 3: if  $(\text{Support} < \mu)$  goto step 6
Step 4:  $S = \text{Hypothesis} / (\text{Total number of rows})$ 
Step 5:  $C = \text{Conclusion} / (\text{Total number of rows})$ 
Step 6: Stop
Step 7: Compare logarithm of odds of happening of heart disease to linear function
Step 8: Compute Probability
Step 9: Stop

```

### C. Explanation

The dataset for this research is taken from UCI machine learning repository, and the dataset name is Statlog for heart disease prediction.

1) **Attributes:** The dataset contained 13 different attributes from which the result has to be predicted. There are exactly 270 cell values associated with each attributes. These



attributes are scanned at first and then minimum support is estimated by observing the dataset.

2) **Support greater than Minimum Support:** If the support is greater than minimum support then confidence and support is calculated for the given dataset and the result is predicted.

3) **Support less than Minimum Support:** In this case, the Joint estimated probability of the given data is computed. And based on those computed probability the results are again predicted.

#### D. Flow Chart

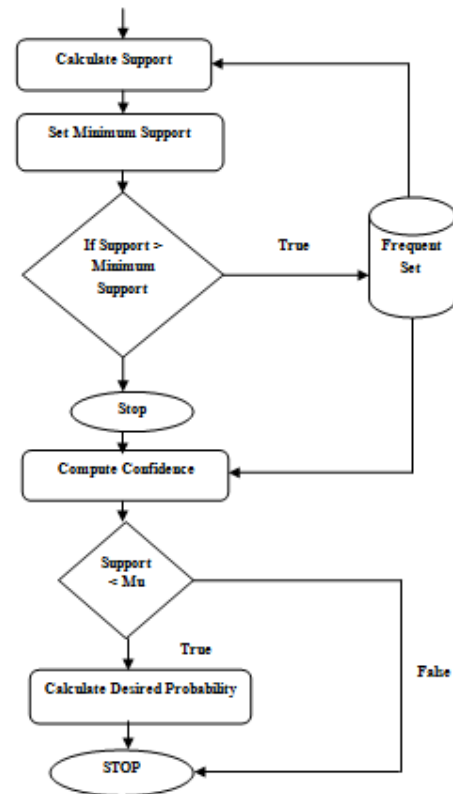
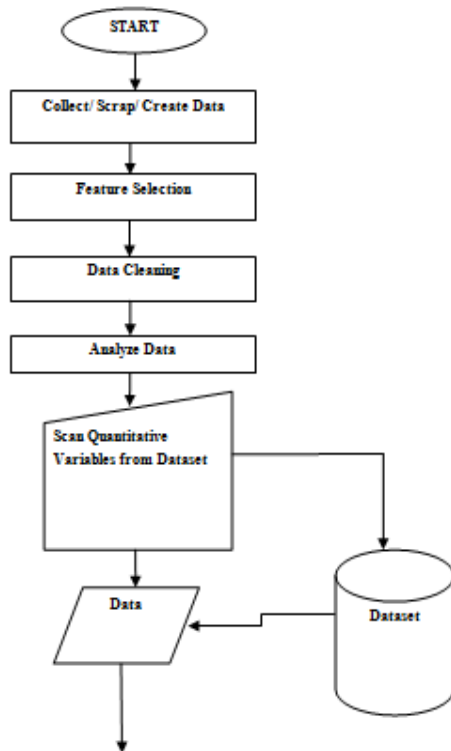


Fig. 2: Flow Chart of Proposed Model

### III. RESULTS

To test the accuracy of the system, 500 different test cases were randomly chosen, out 500 test samples 50 test samples were taken from the original test cases themselves.

To contrast the efficiency of the proposed system, 5 best algorithms were taken namely, Associative Classification and Hybrid Feature Subset Selection, Naïve Bayes, C4.5, k-Nearest Neighbor, and Artificial Neural Network. And the proposed methodology turns out to be more efficient than others. The comparison is shown in the table below:

Table 1: Comparison Table

ACHFSS	Naïve Bayes	C 4.5	KNN	ANN	Proposed Method
95	65	77.5	97.4	94	97.6

The comparison of the proposed system is based on their accuracy for 500 data items taken from different data sources. The formula used for the evaluation for the accuracy is given by:

$$\text{Accuracy} = \frac{(\text{Number of correct prediction} + \text{Number of incorrect prediction})}{(\text{Total number of cases})}$$

The bar chart representation of the comparison is shown the figure below:

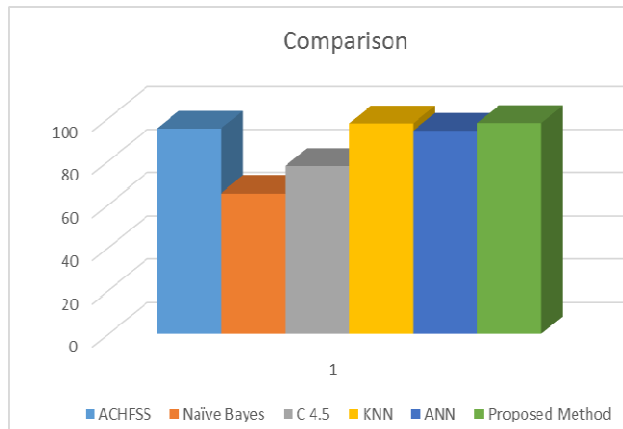


Fig. 3: Comparison

#### IV. CONCLUSION

The target work is to anticipate the risk score of coronary illness from the data with high exactness. We utilized upgraded affiliated principle to focus the presence of affliction in a man. The significant issue with affiliated order is that it works better for clear cut antecedent yet for our situation a large portion of the forerunner present in the data source was quantitative in nature. Furthermore, Associative rule mining, works best for the cases where no example or test outcome are available on the grounds that associative rule mining is an unsupervised learning calculation. What's more, it is constantly shrewd to pick directed learning calculation to foresee scores if some specimen result set are accessible. Yet, administered learning calculation accompanies a punishing expense, and that it is exceptionally time wasteful and

computational costly and if legitimate seeds and weights are not balanced they may not even unite, consequently they require expansive measure of information set and mount focuses ought to be picked by a specialist to make the procedure meet appropriately. Accordingly in this theory associative rule is improved by changing it into a force managed learning calculation. Thusly, now it yields preferred result over some other calculation exhibit in the writing without losing its sharp precision and expedient conduct. The proposed strategy can be utilized as a part of different other disciplinary work to acquire precision with no speed punishment.

#### V. FUTURE SCOPE

As the Healthcare area is dynamic and this issue is a test to the information mining. It is additionally a driving inspiration to the information mining applications in human services. This dynamism offers approach to new skylines and more information mining applications will be utilized to find new examples and affiliations. In the perspective of the subjects analyzed in this study, future information mining studies appear to happen and not restricted but rather in impressive weight, in circulated information mining applications and content mining calculations. With the assistance of information mining calculations, the grouping execution increments. This can be further improved and extended with more expectation calculation for significant life debilitating infections. In this thesis, we have improved the technique for successful heart assault forecast framework utilizing information mining and the enhanced associative rule mining which gives more precise result. It is desktop application where running on stand alone in desktop framework. In future work, we mean to apply on superior customer server or parallel architectures and optical neural system as a classifier model. Utilizing information mining strategies to help human services experts in diagnosing and giving suitable medications to coronary illness and Continuous information can likewise be utilized rather than simply unmitigated information. The customer server application introduced that demand and gets data over the

system so it would simple to get to this application to all clients.

## **ACKNOWLEDGMENT**

I am highly grateful to my CSE department and my college for the great support regarding paper. I really thankful to all my teachers and my guide Ms. Poonamdeep Kaur and co- guide Ms. Lovepreet Kaur who guided me in publishing paper.

## **REFERENCES**

- [1] V. V. R. Bangaru Veera Balaji, "Improved Classification Based Association Rule Mining," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, p. 11, May 2013.
- [2] T. K. and P. S. Chobe, "An Overview of Association Rule Mining Algorithms," *International Journal of Computer Science and Technologies*, vol. 5, p. 5, July 2014.
- [3] D. J. Hand, "Principles of data mining," pp. 621-622.
- [4] N. L. Komal Hadvani, "A Review On Early Heart Diseases Prediction Using Data Mining Techniques," *International Journal of Innovative Research in Technology*, vol. 1, p. 6, August 2014.
- [5] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2011.
- [6] V. S. Nikita Jain "Data Mining Techniques: A Survey Paper," *International Journal of Research in Engineering and Technology*, vol. 2, p. 5, November 2013.
- [7] R. C. and V. Seenivasagam, "Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques," *ICTACT Journal On Soft Computing*, vol. 3, p. 6, July 2013.

# Enhancing Security in Wireless Sensor Network Using Load Balanced Data Aggregation Tree Approach

A.Senthilkumar<sup>1</sup>, K. Madhurabhasini<sup>2</sup>

<sup>1</sup>(Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu), Thanjavur.)

<sup>2</sup>(Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.)

\*\*\*\*\*

## Abstract:

Trust is an important factor in transferring data from the source to destination in wireless AdHoc network. If any node un trust in the transfer the data, the Dynamic Source Protocol calculates the alternate path. Currently, the Dynamic Source Protocol does not have any built-in functionality to calculate an alternate path if the path has a malicious node. Intruder detection system can detect untrust worthy node. However, intruder detection system is very expensive for AdHoc networks and there is no guarantee in detecting a untrust node. In the current research, a trust-based approach is recommended to minimize the overheads of intruder detection system and detect the abnormal behaviour nodes. The data can be send and receive through set the path using the level based scheme to efficiently send the data to the receiver and the data rate can be increased and set the different path to send the data.

**Keywords — Adhoc,propagation.**

\*\*\*\*\*

## I. INTRODUCTION

Wireless Networks with scheduled intermittent connectivity, vehicular that disseminate location-dependent information and pocket-switched networks that allow node to communicate without network infrastructure, are highly partitioned networks that may suffer from frequent disconnectivity. In node transmission, the messages are also named bundles, can be sent over an existing link and buffered at the next hop until the next link in the path appears (e.g., a new node moves into the range or an existing one wakes up). This message propagation process is usually referred to as the node trust strategy, and the routing is decided in an “opportunistic” fashion. In adhoc network a node could misbehave intentionally even when it has the capability to forward the data. Routing misbehaviour can be caused by selfish (or rational) nodes that try to maximize their own benefits by enjoying the services provided by DTN while refusing to forward the bundles for others, or untrust nodes that drop packets or modifying the packets to launch attacks. The recent researches show that routing

misbehaviour will significantly reduce the packet delivery rate and, thus, pose a serious threat against the network performance. Mitigating routing misbehaviour has been well studied in traditional ad hoc networks. These works use neighbourhood monitoring or destination acknowledgement to detect packet dropping, and exploit credit-based and reputation-based incentive schemes to stimulate rational nodes or revocation schemes to revoke malicious nodes.

Even though the existing misbehaviour detection schemes work well for the traditional wireless networks, the unique network characteristics including lack of contemporaneous path, high variation in network conditions, difficulty to predict mobility patterns, and long feedback delay have made the neighbourhood monitoring based misbehaviour detection scheme unsuitable for node. Since there may be no neighbouring nodes at the moment of the misbehaviour cannot be detected due to lack of witness, which renders the monitoring-based misbehaviour detection less. In network, clustering is used as an effective technique to achieve scalability, self-organization, power saving, channel

access, routing etc. Lifetime of sensor nodes determines the lifetime of the network and is crucial for the sensing capability. Clustering is the key technique used to extend the lifetime of a network. Clustering can be used for load balancing to extend the lifetime of a sensor network by reducing energy consumption. Load balancing using clustering can also increase network scalability. Network with the nodes with different energy levels can prolong the network lifetime of the network and also its reliability.

In this existing system the individual user data can be exchanged over the third party nodes. Individual data can be accessed through the third party server, and it can be outsourced. Before outsourcing, the secrecy data to be encrypted and outsource the data. In this system, the particular secrecy data can be maintained by the central authority (CA) to the trust management on behalf of node trust. In this system, the untrust behaviors which may lead to the exposure of the secrecy data. In Existing the access policy based mechanism is not used. The nodes are trusted blindly.

#### **DISADVANTAGES:**

- ❖ In this system, for the individual user having the central authority for the data transmission. Data unsafe.
- ❖ The Data can be accessed by the third party nodes and can be accessed by unauthorized users.
- ❖ Easily Compromised nodes and Reveals Secure Data.
- ❖ The sensitive applications demand secure transmission at the time of deployment of nodes.

#### **PROPOSED SYSTEM:**

In the proposed system, the secure sharing of secrecy data is storing on the trusted nodes in presence of Level based scheme by users. It can be protected using the trusted nodes and level based scheme can be used to trust the particular user data node as per the user needs. These works use neighbourhood monitoring or destination acknowledgement to detect packet dropping, and exploit credit-based and reputation-based incentive schemes to motivate rational nodes or revocation schemes to revoke malicious nodes. In this to

improve security the user is categorized trusted node can be categorized. The data and increased the data rate and the net rate efficient schedule the data and the share the data efficiently.

#### **ADVANTAGES:**

- ❖ Good chance of bringing the source in contact with destination to nodes.
- ❖ High probability of message delivery to succeed.
- ❖ The source and destination come in contact with each other directly.
- ❖ Possible when the source and destination are one hop apart or immediate neighbor of each other.
- ❖ No global or local knowledge about network.

#### **Pseudo code:**

Step 1: Initialize the No of Levels and no of Nodes to construct a WSN network in Data Aggregation Tree

Step 2: Generate the no of levels with no of child in the tree

Step 3: Assign Level Key to each Level of the tree

Step 4: Assign each node key to each level in the tree.

Step 5: Each node can communicate from one level to another using Iterative Filtering.

Step 6: The Source node send info to Destination node by level key and node info key

Step 7: the Key gets updated in Hash Table.

Step 8: The recursive IF is used for secure communication

#### **NETWORK FORMATION**

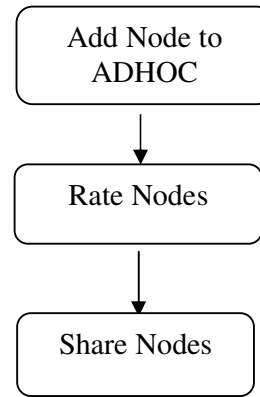
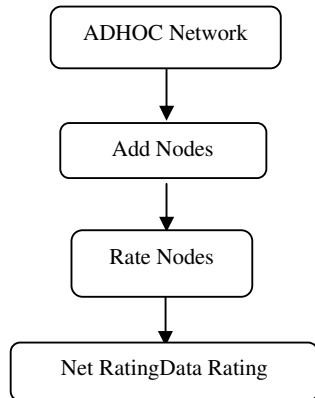
Challenging your Neighbor is defined to trust and authenticate a new node, you can challenge your neighbor to add that node into the network. A node having its neighbors in its friend list does not need to challenge them before a data session.

**Rate nodes:** Initially each node has only those nodes in their friend list that completed the challenge successfully. Sharing of nodes is done in the Share nodes stage as the relation is transitive in nature that relative node of includes in his node list.

**Data rating:** The data rating is updated by a node for its friend on the basis of amount of data it transfers for it.



**Net rating:** FR represents the opinion of the friends of a particular node towards the integrity of another node, towards the integrity of another node, while DR represents a personal opinion of a node derived on the basis of previous data sessions. Both these ratings are important as certain nodes could be selectively malicious based on the holistic metric called as the Net Rating.



### Ensuring Security in Data Sharing Using level based scheme:

ECC is the most secured and advanced cryptography algorithm used for security in data sharing. The ECC algorithm can use a key as small as possible is the main advantage. It helps to protect data sharing in each node of routing. The most common problem occurs in ECC is discrete logarithmic problem, which is used to increase complexity for attackers. Therefore the DLP is combined with ECC cryptography algorithm to ensure security in name of level based scheme.

### Share the Trust Nodes

Nodes sharing is a periodic process which is chiefly responsible for the security of the algorithm. To accomplish friend sharing we use the control packet FREQ (Friend sharing request). The node receiving the FREQ replies with the nodes in its friend list, unauthenticated list and the question mark list.

The rules for friend sharing are as follows.

- Any node can ask for a friend sharing request.
- After friend sharing, challenges are initiated for those nodes which were not in the friend list.
- If a node is already in the friend list the node updates its friend list.

After the friend sharing process is complete a node may start a data session or may sit idle.

### Routing Efficiently

The data can be sent through the single routing path the data sending and receiving time should be increased and the data rate can be decreased and data sharing performance can be delayed by the present using algorithms. The data can be sent and received through set the path using the SEGPSR algorithm to efficiently send the data to the receiver and the data rate can be increased and set the different path to send the data and increased the data rate and the net rate efficient schedule the data and the share the data efficiently.

### 5. CONCLUSION

In this paper the result from simulation and comparison of various routing protocols such as First Contact and Direct Delivery. The data transmission between one node to another node using secure data transmission. The result shows that when we need to achieve higher delivery ratio it will increase the overhead ratio when numbers of nodes are increased. It requires more buffer space to

replicate messages copies. When we replicate more copies it will achieve better delivery ratio but also requires much buffer space to store messages.

## **6. FUTURE ENHANCEMENT**

The control technique for multilevel power converters can be further simplified and generalized to different levels and other class of power converters and inverters. The levels of multilevel configuration can be increased and further improvements in terms of performance and power quality. Data transmission, to avoid such thread, the nodes in the network are monitored by Trusted Authority and set a probabilistic value, the probabilistic value denotes the node trust. So the Probabilistic misbehavior Scheme is used for secure data transmission.

## **REFERENCES**

[1]B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc.EMNLP, Philadelphia, PA, USA, 2002, pp. 79–86.

[2]B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in Proc. HLT-NAACL, New York, NY, USA, 2007, pp. 300–307.

[3]L. Tao, Z. Yi, and V. Sindhwani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge," in Proc. ACL/AFNLP, Singapore, 2009, pp. 244–252.

[4]T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proc. HLT/EMNLP, Vancouver, BC, Canada, 2005, pp. 347–354.

[5]J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua, "Aspect ranking: Identifying important product aspects from online consumer reviews," in Proc. ACL, Portland, OR, USA, 2011, pp. 1496–1505.

[6] Bo Pang<sup>1</sup> and Lillian Lee<sup>2</sup>Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–13

[7] S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 577–588, 2012.c\_Springer-Verlag Berlin Heidelberg (2012)

[8] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. (2010).SentiWordNet 3.0: An enhanced Lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pages 2200–2204, Valletta.

[9] Kessler, B., Nunberg, G., and Schutze, H. 1997. Automatic Detection of Text Genre. In Proc. of 35th ACL/8th EACL.

# Online Shopping Product Aspect and Ranking Using Support Vector Machine Algorithm

<sup>1</sup>R. Bharathi

<sup>1</sup>(Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.)

\*\*\*\*\*

## Abstract:

The peoples are before the purchasing invention to see the product reviews on internet. But some time the reviews are often not confidentiality and provide difficult about product aspect and people could not identify the review information via internet.

Nowadays online shopping plays an excellent role in our life. Peoples are more comfortable to buy online products at the same time manufacturers also provide reliable products .Most retail Websites promotes consumers to write their feedbacks about products to express their opinions on various *aspects* of the products. The web contains outstanding source of consumer opinions. A product may have thousands of aspects. Different kinds of users give their different kinds of opinions .so the volume of the textual information increased rapidly. It is difficult for users to read all the reviews to make a good decision. It is also difficult for manufacturers and providers. These needs extracted aspects and estimated ratings clearly provide more detailed information of users to make decisions and for suppliers to monitor their customers. In this research, we aim to mine and to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because we only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. We do not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Our task is performed in three steps.

(1) Mining product features that have been commented on by customers.

(2)Identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative.

(3) Summarizing the results.

**Keywords — cloud aspects,product opinion.**

\*\*\*\*\*

## I. INTRODUCTION

Online shopping sometimes known as e-tail from "electronic retail" or e-shopping is a form of electronic commerce which allows consumers to directly buy goods or services from a seller over the Internet using a web browser. Alternative names are: e-web-store, e-shop, e-store, Internet shop, web-shop, web-store, online store, online storefront and virtual store. Mobile commerce or m-commerce describes purchasing from an online retailer's mobile optimized online site. The product purchasing in online is one of the part in human life day to day time. Economists have theorized that e-

commerce ought to lead to intensified price competition, as it increases consumers' ability to gather information about products and prices. Research by four economists at the University of Chicago has found that the growth of online shopping has also affected industry structure in two areas that have seen significant growth in e-commerce, bookshops and travel agencies. Generally, larger firms are able to use economies of scale and offer lower prices. The lone exception to this pattern has been the very smallest category of bookseller, shops with between one and four employees, which appear to have withstood the trend. Depending on the category, e-commerce may

shift the switching costs procedural, relational, and financial experienced by customers.

### **1.1.Existing System**

The existing of system is to identify the aspect of product in supervised way that means listening about the product and identify the aspect is manually or not able to identify the aspect is imprecision. The system represent the important aspect of product is imprecision in previously reviews. So this is also impractical the people manually identify the aspect of product from various consumers.

### **1.2. Disadvantages**

- The product aspect identification is imprecision.
- The reviews are not accurately.
- Efficiency of review is low.

## **2. PROPOSED SYSTEM**

The proposed system is SVM ranking is automatically identify the aspect of product from various consumer. The proposed method using significantly perform the product aspect identify and maximize the reviews of product is precision using support vector machine algorithm. The algorithm use to collect the important aspect is simultaneously on aspect frequency and influence important consumer opinion given each aspect to over their overall opinion.

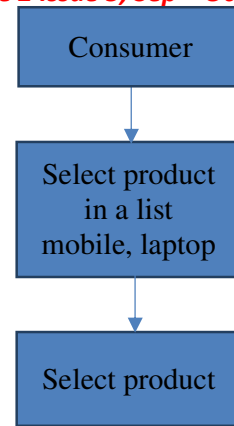
### **2.1. Advantages**

- Easily identify the aspect of the product.
- Increase the efficiency of product review information.
- The product review is accurate.

## **3. METHODOLOGIES**

### **3.1. Consumer choose of product in a categories**

The product in categories is vital role in people life. The consumer choose product in a list of categories via the reviews of the various consumer used on that product. The all type of products are available on list of categories such as mobile, laptop, etc. So the Consumers need the requirement of product purchasing in online via internet in a secure way.



### **3.2. Identify aspect of Product Company**

The consumer first identifies the aspect of the product before the purchasing of the product. The aspect denotes the feature of the product based on company such, amazons, flip kart. The module identifies the aspect of the product in shallow dependency parser. The aspect identified in a way is important aspect is commented by large number of consumer.

### **3.3. Get overall opinion of product**

The people get important aspect of product and their overall opinion of product aspect from already purchased consumers of product. This way help to people purchase a product in efficient manner. The module collects overall opinion of product aspect from numerous consumers in sentimental classifier.

### **3.4. Predict review information based on SVM**

The people first identify the aspect of the product and collect the important aspect from numerous consumers and their overall opinion via using shallow dependency parser and sentiment classifier ranking using SVM Support vector Machine algorithm. Finally the algorithm helps to calculate the aspect information from numerous consumers to provide review and rating of product to help people purchase good product via online.

## **4. OPTIMIZATION OF CROSS DOMAIN SENTIMENT ANALYSIS USING SENTIWORDNET**

The task of sentiment analysis of reviews is carried out using manually built automatically generated lexicon resources of their own with which terms are matched with lexicon to compute the term count for positive and negative polarity.

On the other hand the Sentiwordnet, which is quite different from other lexicon resources that gives scores weights of the positive and negative polarity for each word. The polarity of a word namely positive, negative and neutral have the score ranging between 0 and 1 indicates the strength weight of the word with that sentiment orientation. In this paper, we show that using the Sentiwordnet, how we could enhance the performance of the classification at both sentence and document level.

## **5. CONCLUSION**

Review is to identify the important aspects of a product from online consumer reviews. Our assumption is that the important aspects of a product should be the aspects that are frequently commented by consumers and consumers' opinions on the important aspects greatly influence their overall opinions on the product. Based on this assumption, we have developed an SVM algorithm to identify the important aspects by simultaneously considering the aspect frequency and the influence of consumers' opinions given to each aspect on their overall opinions. We have conducted experiments on 4 popular products in four domains. Experimental results have demonstrated the effectiveness of our approach on important aspects identification.

## **6. FUTURE ENHANCEMENT**

In terms of future scope, a variety of data mining techniques can be used by researchers to simplify customer perceptions and attitudes. Every day, every hour and every minute, tera-bytes of data gets generated from millions of shoppers, yet, retail managers business executives always grapple with relevant information that can help retailers researchers design strategies to generate customer loyalty. Thus data mining can not only be applied in retailing but also can be applied in the other sectors such as banking, medicine, education, and tourism, insurance and so on. Data mining is the task of finding useful information knowledge from huge volume of data. Data mining can be applied through a variety of other techniques such as concept description, cluster analysis, factor analysis, classification and prediction, association analysis, evolution analysis, outlier analysis and many other different tools

## **7. REFERENCES**

- [1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. EMNLP*, Philadelphia, PA, USA, 2002, pp. 79–86.
- [2] B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in *Proc. HLT-NAACL*, New York, NY, USA, 2007, pp. 300–307.
- [3] L. Tao, Z. Yi, and V. Sindhvani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge," in *Proc. ACL/AFNLP*, Singapore, 2009, pp. 244–252.
- [4] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. HLT/EMNLP*, Vancouver, BC, Canada, 2005, pp. 347–354.
- [5] J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua, "Aspect ranking: Identifying important product aspects from online consumer reviews," in *Proc. ACL*, Portland, OR, USA, 2011, pp. 1496–1505.
- [6] Bo Pang<sup>1</sup> and Lillian Lee<sup>2</sup> Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–13
- [7] S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 577–588, 2012.c\_Springer-Verlag Berlin Heidelberg (2012)
- [8] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. (2010). SentiWordNet 3.0: An enhanced Lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pages 2200–2204, Valletta.
- [9] Kessler, B., Nunberg, G., and Schutze, H. 1997. Automatic Detection of Text Genre. In Proc. of 35th ACL/8th EACL.
- [10] Esuli A, Sebastiani F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings from International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.
- [11] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in *Proc. IT&T Conf.*, Dublin, Ireland, 2009.



# Resilient Key Sharing Approach Based on Multimode Authentication Scheme

<sup>1</sup>A.Senthilkumar, <sup>2</sup>R.Divya

<sup>1</sup>Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu), Thanjavur.

<sup>2</sup>Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.

\*\*\*\*\*

## Abstract:

The history of Information Security begins with the study of computer security. Security concern arises from various parameters to safeguard physical location, hardware and software from the outside threats. The research work carries out essential security needs and adoptions of security policies based on Resilient Key Distribution Mechanisms (RKDM) to safeguard information transmissions over networks. The adoption of sender and receiver whom are called as 'clients' should register themselves initially. The registration work of them clients are maintained in separate databases maintained for them, and whenever each user makes a login request, thereby an individual private key is created by the clients themselves. To know their own authentication, a centralized server which is called as an 'Authentication Server (AS)' is maintained to monitor the data transaction of the clients and the application which runs between them. After confirming the registration, the server issues resilient key (Public keys) for each of the user for their data transactions in the networks. The authentication server issues resilient key only after verification of each client who register themselves within it.

**Keywords — Information security, Resilient key distribution mechanisms(RKDM) private key, authentication ,authentication server, resilient key(public keys) ,RSA algorithm ,object set , user set, node awareness ,multimode.**

\*\*\*\*\*

## I. INTRODUCTION

Authentication is referred as 'User identity who claims any resources in the network'. Authentication plays a key role in preventing unauthorized and corrupted messages to safe the data transmission in wireless networks (WN). For this reason, many authentication schemes have been proposed in literature to provide message authenticity and integrity verification for wireless networks (WNs). Message authentication schemes can largely be divided into two categories: public-key based approaches and symmetric-key based approaches. The symmetric-key based approach requires complex key management, lacks of scalability, and is not resilient to large numbers of node compromise attacks since the message sender and the receiver have to share a secret key. An intruder can compromise the key by capturing a single sensor node. In addition, this method does

not work in multicast networks. The idea of this scheme is to implement resilient key sharing based on RSA algorithm which considers secret sharing, where the key sizes are determined initially from 1024 bits.

This approach offers information-theoretic security of the shared secret key when the number of messages transmitted as specified by the algorithm key size. The intermediate nodes verify the authenticity of the message through an evaluation. The message is transmitted along with the Server issued resilient keys. Every nodes in the network can share files to other nodes only after verifying their identities. The Research proposal includes registration module for user registrations, authentication module to verify the users and key sharing and file sharing and secured module where the algorithm is implemented and the node status to identify the awareness of the user which narrates the user, object , the file and all the user's

authentication. The file shared to node and verify the keys and generate the resilient key. The number of local hosts connected in this work remains limited and can be scaled according to the organization or network in future. While achieving compromise resiliency, flexible-time authentication and source identity protection.. Both theoretical analysis and simulation results demonstrate that our proposed scheme is more efficient than the RSA algorithms under comparable security levels. The advancement of wireless communication technologies and rooted computing, are being widely adapted into many applications through networks and many active researches on related subject are being carried out. Several nodes are connected to build the wireless networks. The large numbers of devices are the interconnected to form a network. The WN make use of a number of nodes within or neighboring the area of event to not only collect and integrate but also process and relay the information. The . registration of clients or data transaction but also confirms one of the security policies named Authentication.

### **PROBLEM DESCRIPTION**

#### **Existing system:**

"The existing system based on RSA has not fixed network limitation ".the symmetric-key based approach needs complex key management, lacks of scalability, and is not resilient to large numbers of node compromise attacks since the message sender and the receiver have to share a secret key. The sender generate message Authentication Code (MAC) for each transmitted message by using the shared key. However, for this basis, the authenticity and integrity of the message can only be checked by the node with the secret key, which is usually shared by a group of nodes. The existing system provides the use of RSA algorithm with not specifying the user awareness status, which is difficult for any user to locate the identity this research proposal solves the issues by Identifying the user , node of the file to be kept secret. For a public-key based approach, each message is transmitted along with the digital signature of the message which is generated using the sender's private-key. Every intermediate forwarder and the final receiver can authenticate the message using the sender's public-key. The drawbacks of the

public-key based approach is the high computational overhead.

### **PROBLEM DEFINITION**

User authentication plays a key role in preventing unauthorized and corrupted messages to safe the data transmission in wireless networks (WN). For this reason, many authentication schemes have been proposed in literature to provide message authenticity and integrity verification for wireless networks (WNs). user authentication schemes can largely be divided into two categories: public-key based approaches and symmetric-key based approaches. In this research proposal on a server called a "Resilient key".

The symmetric-key based approach requires complex key management, lacks of scalability, and is not resilient to large numbers of node compromise attacks since the message sender and the receiver have to share a secret key. The shared key is used by the sender to generate a message authentication code for each transmitted message. The authenticity and integrity of the message can only be verified by the node with the shared secret key, which is generally shared by a group of nodes. . The here public key server as authentication server, issues resilient keysecret RSA based message authentication scheme. The idea of this scheme is similar to secret sharing, where the is determined by the degree of the RSA This approach offers information-theoretic security of the shared secret key when the number of messages transmitted by user. The intermediate nodes verify the authenticity of the message through a RSA evaluation.

### **3.2 PROPOSED SYSTEM**

For the public-key based approach, each message is transmitted along with the digital signature of the message generated using the sender's private key. Every intermediate forwarder and the final receiver can authenticate the message using the sender's public key. One of the limitations of the public-key based scheme is the high computational overhead. The recent progress on RSA algorithm shows that the public key schemes can be more advantageous in terms of computational complexity, memory usage, and security resilience, since public-key based approaches have a simple and clean key

management. Our scheme enables the intermediate nodes to authenticate the message so on. In this work are used five modules that can be used: Registration, authentication, file sharing, node status, and last one performance analysis. Registration module suggests the client or node based on their own login name or password. Authentication server checks the secret key of the user name, file sharing: this module file or any application is chosen it going to be shared. Node status: the node in idle or busy state or secured. Both theoretical analysis and simulation results demonstrate that our proposed scheme is more efficient than the RSA algorithms used for encryption, decryption security levels.

## METHODOLOGY

### Design steps:

1. Let the nodes be assumed as  $n_1, n_2, n_3$
2. Server is notated as 'AS1'
3. Given the nodes initialize then according to their private key assumptions say  $ln_1, pw_1$  for node ' $n_1$ ',  $ln_2, pw_2$  for node ' $n_2$ ',  $ln_3, pw_3$  for node ' $n_3$ '.
4. Repeat the initialization procedure if added up clients are there in network.
5. Call the authentication procedure for the initial nodes
6. Here AS1 verifies for authentication by the nodes private key  $ln_1, ln_2$  for all clients.
7. Authentication server rejects the invalid node if they do not possess secret keys
8. Call resilient key authentic procedure, let the resilient key be  $R_1, R_2, \dots, R_n$
9. Allocate resilient key  $r_1$  for  $n_1$ ,  $r_2$  for  $n_2$  and  $r_3$  for  $n_3$ ;
10. Allocate until the resilient keys of authentic server ends up
11. Call the file sharing procedure for security implementations
12. Let the file, or node or application to run on node 1 be initialized say  $f_1 = 1$ ,  $a_1 = 1$ ,  $n_1 = 1, n_2 = 1$  and so on
13. Let the file  $n_1 = f_1$  (file  $f_1$  is in use) by the node  $n_1$
14. Call the security procedure call  $sec()$ ;
15. Calculate the private key & public key computations.
16. Call the encrypt procedure if file is in node 1 or application is in node 1
17. Call the decrypt procedure and return back to produce plaintext
18. Check the node status procedure to print the usernames, network boundaries, key sizes and other too
19. Call the perform analysis procedure for all files, user and applications
20. Thus multimode authentication schemes are utilized in this research work.

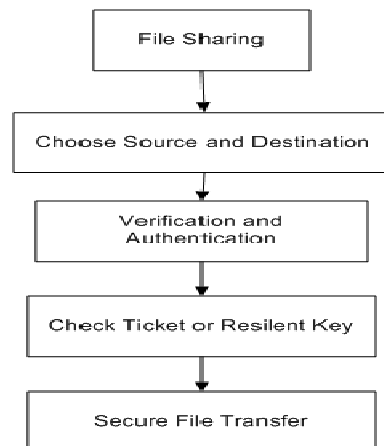
### 1) Registration:

Registration module suggests the clients or node registration based on their own login name or password. Initially it occurs for a single user and for a single receiver and updates to more no. of users at the end. The login name or password generated is assumed to be their secret key.

### 2) Authentication Module:

In the authentication module, the authentication server checks the secret key of the user namely (l name, pw) and validates the user. This can be a repeated process for additional clients also say node2 or node3 and so on.

In the authentication module (Sub module) as another part, the server provides or issues Resilient keys which is the Public key to the registered Users. The server database maintains all the **registered users who are validated** based on their private key (l name, pw) and the public key (resilient key1...resilient key2) up to the node limit.



### 3)File Sharing:

In this module, the file or any application say f1 or a1 is chosen. If it is going to be shared, the file must be secured USING RSA Algorithm before sharing the file, the clients records, RSA algorithm the security for authentication purpose.

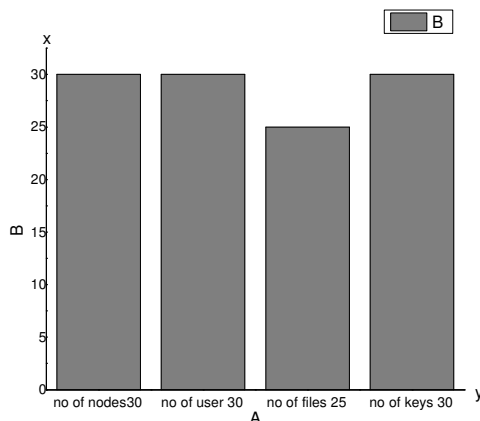
For example, if a content or a word is chosed in a sample file, it must be encrypted using RSA formula,  $c = m^e \bmod n$  where m is the plaintext, e to the power and must be decrypted  $m = c^d \bmod n$ .

### 4)Node Status:

The RSA algorithm proposes the study of network security but it does not analyzed the study of network limitation, but the research work proposed suggests the network boundary since nodes or clients must be framed in a boundary to achieve information security very fastly. This can be scalable in future according to the organizational needs. In this module, the work checks which user say u1, is viewing which file, say f1 and up to which network boundary say  $u_n$ , is in what status [ the node is in idle state or busy state or secured].

### 5)Results Analysis and Performance Efficiency:

Performance analysis module suggests the research work carried to how much extent or an average the user or file or nodes got secured. For this the nodes can be arranged in a linear fashion next to next. The files secured based on the keys must be matched up with the nodes. The nodes with file must be equally secured proportionally to the maximum limit. For example if node1 checks 5 files means (1: 5), secondly node 2 checks 10 files( 1:10) secured or the contents secured.



## 7. CONCLUSION

In this research proposal analysis the total number of client with the RSA algorithm implement the secure files. In we used five modules which initially login the client of the look successfully concludes the concept Of authentication principle in a elaborated manner of an is provided in a multimode fashion. User authentication scheme. improve the security in wireless Networks. An efficient Source anonymous message authentication schemeAn intermediate nodes are authenticate and allow to transmit a message, does not have the network boundary limitation that is unlimited number of messages are secret key verified than server compare based scheme Proposed scheme is more efficient than the RSA-based scheme such as memory and security .

## 8. FUTURE ENHANCEMENT

The number of nodes at the initial stage can be limited. in future, the nodes of the research proposal can be extended to maximum number of nodes.

Security analysis showed that the proposed protocol fulfils the security benefits that a secure user authentication scheme should provide and can resist to various possible attacks such as SSL/TLS man-in-the-middle attack, offline dictionary and brute force attack, and message modification or insertion attack. Future scope in this work is to investigate the possibility of using cloud computing technology to improve the portability and recovery of our scheme. We also plan to provide a secure design for user's session management. While the proposed scheme needs public key cryptography, then we will further focus on elliptic curve cryptography which offers faster computations and less power use. The application uses can be enlarged as per any organization wish with regard to number of nodes increased to scale in the future.

## 7.REFERENCES

- [1]B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc.EMNLP*, Philadelphia, PA, USA, 2002, pp. 79–86.
- [2]B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in *Proc.*

*HLT-NAACL*, New York, NY, USA, 2007, pp. 300–307.

[3] L. Tao, Z. Yi, and V. Sindhwani, “A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge,” in *Proc. ACL/AFNLP*, Singapore, 2009, pp. 244–252.

[4] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proc. HLT/EMNLP*, Vancouver, BC, Canada, 2005, pp. 347–354.

[5] J. Yu, Z.-J. Zha, M. Wang, and T. S. Chua, “Aspect ranking: Identifying important product aspects from online consumer reviews,” in *Proc. ACL*, Portland, OR, USA, 2011, pp. 1496–1505.

[6] Bo Pang<sup>1</sup> and Lillian Lee<sup>2</sup> Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–13

[7] S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 577–588, 2012.c\_Springer-Verlag Berlin Heidelberg (2012)

[8] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. (2010). SentiWordNet 3.0: An enhanced Lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pages 2200–2204, Valletta.

[9] Kessler, B., Nunberg, G., and Schutze, H. 1997. Automatic Detection of Text Genre. In Proc. of 35th ACL/8th EACL.



# Secure Multiparty Computation on Multiple Clouds

<sup>1</sup>K.Ravikumar, <sup>2</sup>S. Thamizharasi

<sup>1</sup>Asst.professor, Dept.of.Computer science, Tamil University (Established by the Govt.of.Tamilnadu), Thanjavur.

<sup>2</sup>Research Scholar, Dept.of.Computer Science, Tamil University, Thanjavur.

\*\*\*\*\*

## Abstract:

Cloud Computing is a model for delivering information technology services in which resources are retrieved from the internet through web-based tools and applications, rather than a direct connection to a server. A distributed cloud storage system contains collection of storage servers which continuously provides storage services to the third party or clients. The distributed cloud storage system must maintain the data confidentiality over the stored data in the storage server. This can be done by encoding over the encrypted data on the storage server. The distributed cloud storage system also maintains the robustness and functionality over the encoded and encrypted data. The distributed cloud storage system support data forwarding operations over encoded and encrypted messages.

**Keywords — Proxy Encryption, Cloud.**

\*\*\*\*\*

## I. INTRODUCTION

Storing data in a third party's cloud system causes serious concern over data confidentiality and there are some functionality restrictions on the storage system. We focus on designing a cloud storage system for data robustness, confidentiality, and improve the functionality of the storage server. These all can be achieved through a threshold proxy re-encryption scheme and integrate it with a decentralized erasure code such that a secure distributed storage system is formulated.

### System setup:

Login/Register  
Key generator (PK and SK)  
Share to Key server

### Data storage:

Storing data in the storage server

### Data forwarding:

Forward data to another user

### Data retrieval:

By data owner  
Received data

## 2.PROPOSED SYSTEM TECHNIQUES

### Proxy Re-Encryption Schemes:

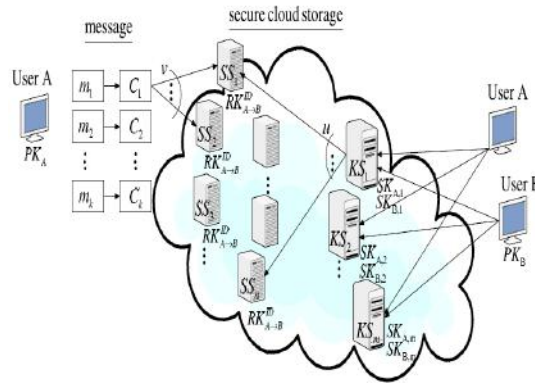
In a proxy re-encryption scheme, a proxy server can transfer a cipher text under a public key PKA to a new one under another public key PKB by using the re-encryption key RKA→B. The server does not know the plaintext during transformation. When user A wants to store messages are first encrypted by the owner and then stored in a storage server. When a user wants to share his messages, he sends a re-encryption key to the storage server. The storage server re-encrypts the encrypted messages for the authorized user. Thus, their system has data confidentiality and supports the data forwarding function.

### Advantages

The storage server can able to transfer the stored user's data into another user by analyzing the stored user ID and our system is distributed storage system it can perform independently.

Our cloud storage system maintains the robustness, confidentiality and functionality.

To store data on the storage server performs encryption and encoding so that it maintains the data confidentiality.



### 3. MODULES DIAGRAM AND DESCRIPTION

#### System setup:

- **Login/Register:**

In Login Form module presents users a form with username and Password fields. If the user enters a valid username/password combination they will be granted to access data. If the user enter invalid username and password that user will be considered as unauthorized user and denied access to that user.

If he is a new user he needs to enter the required data to register the form and the data will be stored in server for future authentication purpose.

#### Key generator (PK and SK)

The key generator generates the public key and secret key for the new user. These public and private or secret keys are used to encrypt and decrypt the messages for data confidential purpose. Usually public key is used to encrypt the data and secret key or private key is used to decrypt the cipher text to get the original plain text.

#### Share to Key server

The user has to share his secret key to randomly chosen key server. This secret key is used to decrypt the encoded message when the authenticated person wants to share his data or retrieve his data.

### 4. DATA STORAGE

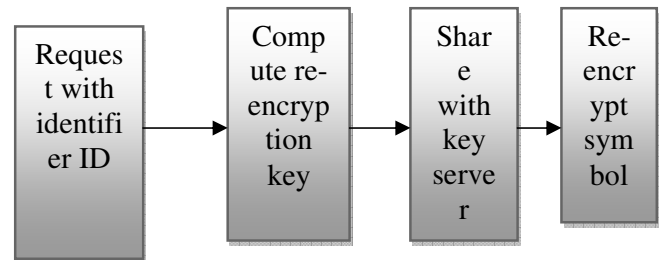
- **Storing data in the storage server**

In the data storage phase, user A encrypts his message M and dispatches it to storage servers.

A message M is decomposed into k blocks  $m_1, m_2, \dots, m_k$  and has an identifier ID. User A encrypts each block  $m_i$  into a cipher text  $C_i$  and sends it to v randomly chosen storage servers. Upon receiving cipher texts from a user, each storage server linearly combines them with randomly chosen coefficients into a codeword symbol and stores it.

### 5. FORWARD DATA TO ANOTHER USER

When user A wants to forwards his encrypted message with an identifier ID stored in storage servers to user B such that B can decrypt the forwarded message by his secret key. To do so, A uses his secret key  $SK_A$  and B's public key  $PK_B$  to compute a re-encryption key  $RK_{A \rightarrow B}^{ID}$  and then sends  $RK_{A \rightarrow B}^{ID}$  to all storage servers. Each storage server uses the re-encryption key to re-encrypt its codeword symbol for later retrieval requests by B. The re-encrypted codeword symbol is the combination of cipher texts under B's public key.



#### By data owner

The first case is that a user A retrieves his own message. When user A wants to retrieve the message with the identifier ID, he informs all key servers with the identity token. A key server first retrieves original codeword symbols from randomly chosen storage servers and then performs partial decryption on every retrieved original codeword symbol. The key server sends the partially decrypted codeword symbols to user A. Then user A applies decryption on collected cipher text to recover the blocks and then combines them to get original data.

- **Received data**

The second case is that a user B retrieves a message forwarded to him. User B sends the request to key server with identifier ID. After authenticating the user B key server decode the re-

encrypted codeword symbol. The key server sends the partially decrypted codeword symbols to user A. Then user A applies decryption on collected cipher text to recover the blocks and then combines them to get original data.

## **6. GIVEN INPUT AND EXPECTED OUTPUT**

### **System setup:**

#### **Login/Register**

Input: User has to give required data to login or register to access the cloud storage

Output: Storage system permits them to access the cloud storage

#### **Key generator (PK and SK)**

Input: User details

Output: Generate the SK and PK

#### **Share to Key server**

Input: SK to the Key server

Output: SK is stored in Key server

#### **Data storage:**

##### **Storing data in the storage server**

Input: Encrypted message to the Storage Server

Output: Encode and store it

#### **Data forwarding:**

##### **Forward data to another user**

Input: Request with identification ID

Output: Apply Re-encryption scheme and store it

#### **Data retrieval:**

##### **By data owner**

Input: Request with Identification ID

Output: Original Message

## **7. CONCLUSION**

The threshold proxy re-encryption scheme supports encoding, forwarding, and partial decryption operations in a distributed way. To decrypt a message of  $k$  blocks that are encrypted and encoded to  $n$  codeword symbols, each key server only has to partially decrypt two codeword symbols in our system. By using the threshold proxy re-encryption scheme, we present a secure cloud storage system that provides secure data storage and secure data forwarding functionality in a decentralized structure. Moreover, each storage

server independently performs encoding and re-encryption and each key server independently perform partial decryption.

## **8. FUTURE ENHANCEMENT**

Our Key server performs the main role in our distributed storage system. This Key server performs the important role key management. But our proposed system doesn't provide any security over this Key Server. The attacker or intruder can attack the key server to get the secret key because there is no security provided to the Secret Key. So as a future work we focus on key server for giving more secure to our storage system. To overcome this problem we are going apply the encoding over the secret key before it store in the Key server. The Key server can decode this encoded Secret Key.

## **9. REFERENCES**

1. A. Adya, W.J. Bolosky, M. Castro, G. Cermak, R. Chaiken, J.R. Douceur, J. Howell, J.R. Lorch, M. Theimer, and R. Wattenhofer, "Farsite: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment," Proc. Fifth Symp. Operating System Design and Implementation (OSDI), pp. 1-14, 2002.
2. M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu, "Plutus: Scalable Secure File Sharing on Untrusted Storage," Proc. Second USENIX Conf. File and Storage Technologies (FAST), pp. 29- 42, 2003.
3. H.-Y. Lin and W.-G. Tzeng, "A Secure Decentralized Erasure Code for Distributed Network Storage," IEEE Trans. Parallel and Distributed Systems, vol. 21, no. 11, pp. 1586-1594, Nov. 2010.
4. R. Bhagwan, K. Tati, Y.-C. Cheng, S. Savage, and G.M. Voelker, "Total Recall: System Support for Automated Availability Management," Proc. First Symp. Networked Systems Design and Implementation (NSDI), pp. 337-350, 2004.
5. A.G. Dimakis, V. Prabhakaran, and K. Ramchandran, "Ubiquitous Access to Distributed Data in Large-Scale Sensor Networks through Decentralized Erasure Codes," Proc. Fourth Int'l Symp. Information Processing in Sensor Networks (IPSN), pp. 111-117, 2005.

6. Q. Tang, "Type-Based Proxy Re-Encryption and Its Construction," Proc. Ninth Int'l Conf. Cryptology in India: Progress in Cryptology (INDOCRYPT), pp. 130-144, 2008.
7. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," Proc. 14th ACM Conf. Computer and Comm. Security (CCS), pp. 598-609, 2007.
8. G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," Proc. Fourth Int'l Conf. Security and Privacy in Comm. Networks (SecureComm), pp. 1-10, 2008.
9. H. Shacham and B. Waters, "Compact Proofs of Retrievability," Proc. 14th Int'l Conf. Theory and Application of Cryptology and Information Security (ASIACRYPT), pp. 90-107, 2008.
10. K.D. Bowers, A. Juels, and A. Oprea, "HAIL: A High-Availability and Integrity Layer for Cloud Storage," Proc. 16th ACM Conf. Computer and Comm. Security (CCS), pp. 187-198, 2009.

# A Survey on Log Mining: A Data Mining Approach for Intrusion Detection

Smita P. Bhapkar<sup>1</sup>, Shubhangi S. Dhamane<sup>2</sup>, Yogita S. Kandekar<sup>3</sup>, Khushbu S. Lodha<sup>4</sup>

<sup>1,2,3,4</sup>(Student of BE computer Chhatrapati College of Engineering, Nepti, Ahmednagar, Savitribai Phule, Pune University)

\*\*\*\*\*

## Abstract:

There are many approaches present in today's world to protect data as well as network. One of the way is an Intrusion Detection System (IDS) to make data more secure. Many researches are done in the field of intrusion detection, but the main concentration of these researches is on the networks and operating system. The unauthorized access may lead to break the integrity of the system as it may be in the form of execution of malicious transaction. E-commerce is one of the sectors suffers from million dollar losses only because of these unauthorized activities and malicious transactions. So, it is today's demand to detect malicious transactions and also to provide some protection. In this paper, we provided the detection system for intrusion detection in the e-commerce system and we are also trying to avoid different types of attack by applying different preventive measures. For detecting malicious transactions, we are going to use one of the data mining algorithm weighted data dependency rule miner for our eCommerce database IDS. Which, extracts the read-write dependency rules to check whether the transactions are malicious or not. This system finds the malicious transactions as well as identify the transactions that performs read write operations without permission.

**Keywords —** Data mining , Database security , Intrusion detection

\*\*\*\*\*

## I. INTRODUCTION

Every organization is associated with more valuable information. Data is important and so it should be consistent and correct. Intrusion Detection System (IDS) is one of the ways to make data more secure. The Intrusion detection technique is a technology used to observe computer activities for finding security violations. When an intrusion takes place computer system is compromised. Intrusion detection is the process of identifying and responding to malicious activity in the various transactions on the internet.

In many systems, firewalls are used for intrusion detection, but they sometimes fail in detecting attacks that take place. To overcome this drawback of firewalls, different data mining techniques are used that handle intrusions occurring from many transactions in databases.

Different data mining techniques like classification, clustering and association rule can be

used for monitoring and analyzing the network traffic and thereby detecting intrusions. [4]

In today's online world, we use the online eCommerce websites for many purposes. So, different E-commerce applications are becoming targets for criminal attacks. Malicious activities are identified from the huge amount of data sets which is generated widely for each transaction in eCommerce system such as logging data and user behavior. The data collected through logs and user behavior, can be a great advantage for intrusion detection to learn from the previous attacks [7]

When making purchases online, consumers should have a general sense of security. Even though there is no way to completely secure consumer information, businesses should take as many precautions as possible, while still allowing for usability. Technology is constantly changing, criminals are constantly finding new methods of attack, and it is the responsibility of users and administrators to use it in a way that is ethical and complies with all laws and regulations. Businesses



need to ensure their e-commerce infrastructures are up-to-date with the latest updates and security necessities [5].

In recent years, data mining has lots of attention in the industry due to the wide availability of the huge volume of unstructured data. Data mining, commonly refers to the process of determining patterns or extracting useful models from large observed data. Recently, researchers have started to use data mining techniques in the system security and especially in intrusion detection systems [8].

Many researchers use different techniques of data mining for detection of intrusion. In this paper, for intrusion detection we are making use of the weighted data dependency rule mining. The read-write dependency rules are used to check whether the coming transaction is malicious or not. This approach mine the dependency among various attributes in a database. The transactions that do not follow these dependencies are called as malicious transactions. After detection of intrusion the appropriate action can be taken.

## **II. LITERATURE SURVEY**

Different approaches have been proposed by researchers to address the problem of identifying malicious database transactions.

[1] Ms. Apashabi Chandkhan Pathan, Mrs. Madhuri A. Potey “Detection of malicious transaction in database using log mining approach ”2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies. In this paper they use log mining approach for detecting anomalous transaction. They define some rules such as data item dependency rules, data sequence rules, domain dependency rules, and domain sequence rules. Database transaction that does not comply these rules are called as malicious transaction.

[2] Ashish Kamra, Elisa Bertino, “Guy mechanisms for Database Intrusion Detection and Response”, Proceedings of the Second SIGMOD PhD Workshop on Innovative Database Research, ACM

2008. The main focus of this paper is to develop higher security solutions for Protecting data reside in database management system (DBMS).The strategy used for detecting intrusion is developing an intrusion detection system within server which is capable of identifying anomalous user request to a DBMS. The idea behind this system is learning profiles of user and application interacting with the database and when the database request deviates from these profiles called as anomalous.

[3] Srivastava A, Sural S, and Majumdar A. K, “Database intrusion detection using weighted sequence mining”, Journal of Computers, IEEE 2006. In this paper intrusion detection system (IDS) is used for detecting potential violations in database security. They propose an algorithm to find the dependency among various data items in a relational database system and the transaction which does not follow this dependency rule is termed as malicious transaction. As every database has some sensitive attribute for malicious modification. The algorithm novel weighted data dependency rule mining is used to detect modification of sensitive attribute. Sensitivity levels of attributes can be captured syntactically while data modeling by using a simple extension of the E-R diagram notations.

[4] Ms. Radhika S. Landge, Mr. Avinash P. Wadhe “Review of Various Intrusion Detection Techniques based on Data mining approach”, International Journal of Engineering Research and Applications (IJERA). Recently data mining techniques are emerging trends in detection of intrusion. In this paper they give various data mining techniques for detecting intrusion such as classification, clustering, association rule mining. These techniques identify intrusion by analyzing network data.

[5] Syed (Shawon) M. Rahman, Ph.D. and Robert Lackey “E-COMMERCE SYSTEMS SECURITY FOR SMALL BUSINESSES” International Journal of Network Security & Its Applications (IJNSA) March 2013. In this paper they discuss how the attacks are carried out and how to secure the network from these attacks with minimum cost. They discuss various protection methods such as biometrics, smart cards, wireless security for

protecting eCommerce system for various types of attacks.

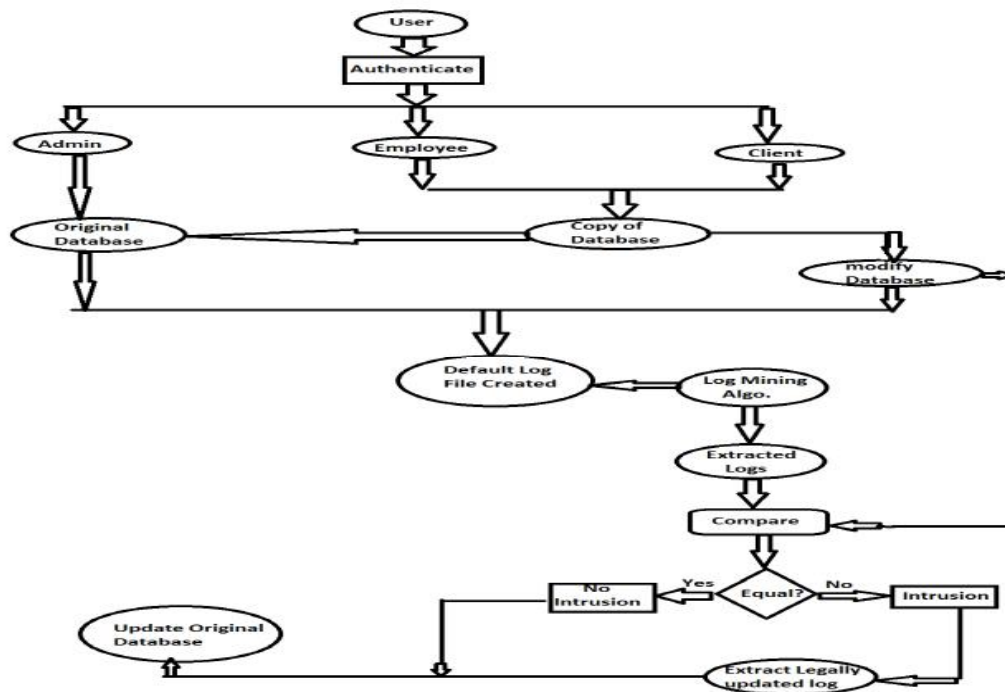
### III. PROPOSED SYSTEM

#### A. Working flow of the system

The fig. is the working flow of the intrusion detection on data mining using log mining approach. The user can be an administrator, an auditor or a third party user (employee). The user will authenticate giving his legal identity. Depending on the type of user the access to the data will be granted. If the user is an administrator, the access to the original database is granted. If the user is an auditor or an employee, a copy of the database is accessed not the original one. All the modifications and updating by the auditor and employee are done only on the copy of the database.

For every change or access to the data, whether on original or on the copy, a default log file is created, which will have all the details of what changes were done, at what time and by whom. Only administrator has the access to the log files.

Using the Log Mining algorithm, the administrator can extract the specific log files, which will help to compare the original database and the modified copy of the database.



**Fig. 1. Schematic representation - Working flow of IDS**

If the changes made are legal or the same as expected, then there's no intrusion in the database. If the changes are made illegal, then intrusion is detected and action can be taken.

For detecting malicious transactions, the data mining algorithm weighted data dependency rule miner is used for our eCommerce database IDS. Which, extracts the read-write dependency rules to check whether the transactions are malicious or not. The transactions which will not follow these dependencies are considered to be a malicious transaction. After detection of intrusion the appropriate action can be taken.

#### **IV. CONCLUSION**

Malicious attacks in e-commerce websites can be easily identified using our proposed system. Our system provided the detection system for the intrusion detection in the e-commerce system. To detect malicious transactions, we are going to use data mining algorithm which uses read-write dependency rules for detecting malicious transactions. If any malicious transaction is detected by our IDS it will be rollback the tampered data with the original. We will use logs files to rollback the data.

This system can be used in various sectors such as Banking, Business, Medical systems for securing the valuable information.

#### **V. REFERENCES**

- [1] Ms. Apashabi Chandkhan Pathan, Mrs. Madhuri A. Potey "Detection of malicious transaction in database using log mining approach "2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies.
- [2] Ashish Kamra, Elisa Bertino, "Guy mechanisms for database intrusion detection and response", Proceedings of the Second SIGMOD PhD Workshop on Innovative Database Research, ACM 2008.
- [3] Srivastava A, Sural S, and Majumdar A. K, "Database intrusion detection using weighted sequence mining", Journal of Computers, IEEE 2006.

- [4] Ms. Radhika S. Landge, Mr. Avinash P. Wadhe "Review of Various Intrusion Detection Techniques based on Data mining approach" International Journal of Engineering Research and Applications (IJERA) June 2013.

- [5] Syed (Shawon) M. Rahman, Ph.D. and Robert Lackey "E-COMMERCE SYSTEMS SECURITY FOR SMALL BUSINESSES" International Journal of Network Security & Its Applications (IJNSA) March 2013.

- [6] Ho, SweeYenn (George) "Intrusion Detection - Systems for today and tomorrow" Version 1.2e, SANS Institute

- [7] Daniel Massa & Raul Valverde, "A Fraud Detection System Based on Anomaly Intrusion Detection Systems for E-Commerce Applications" Computer and Information Science; Vol. 7, No. 2; 2014 Published by Canadian Center of Science and Education

- [8] Abhinav Srivastava<sup>1</sup>, Shamik Sural, and A.K. Majumdar, "Weighted Intra-transactional Rule Mining for Database Intrusion Detection", 2006

- [9] William A. R. Weiss, "An Introduction to Set Theory", October 2, 2008.

- [10] Morten Blomhoj, Thomas Hojgaard Jensen, "Developing Mathematical Modelling Competence: Conceptual Clarification and Educational Planning", July 2003.

#### **VI. BIOGRAPHIES**

**R. Tambe** is currently working as Asst. Professor in Computer Engineering Department, Chhatrapati College of Engineering, Nepti, Ahmednagar and Maharashtra India. His research interest includes data mining, network security.

**Bhaskar Smita** is pursuing B.E Computer Engg. in SCSMCOE, Nepti, Ahmednagar. Her areas of research interests include Database Security and Data mining.

**Dhamane Shubhangi** is pursuing B.E Computer Engg. in SCSMCOE, Nepti, Ahmednagar. Her areas of research interests include Database Security and Data mining.

**Kandekar Yogita** is pursuing B.E Computer Engg. SCSMCOE, Nepti, Ahmednagar. Her areas of research interests include Database Security and Data mining.

**Lodha Khushbu** is pursuing B.E Computer Engg. in SCSMCOE, Nepti, Ahmednagar. Her areas of research interests include Database Security and Data mining.

# Bluetooth Message Hopping Chat Application

Kirti Karande<sup>1</sup>, Ibrahim Shaikh<sup>2</sup>, Tanzeel Shaikh<sup>3</sup>, Hardik Vaghela<sup>4</sup>  
<sup>1,2,3,4,5</sup>(Information Technology, Rajiv Gandhi Institute of Technology, Versova, Mumbai.)

\*\*\*\*\*

## Abstract:

Places where internet and cellular connectivity is weak and mobile devices are used, Bluetooth can prove to be a beneficial communication medium. Also, Android operating systems are quite often used in the organizations by the employees and the staff. Hence, we propose a Bluetooth based Android Chat Application that can communicate with a BT device outside the Bluetooth range of the Source device. For this, an adhoc network is created that acts as a mediator between the source and the destination BT device. The path within this network is found using the MPR flooding technique. As privacy of messages is crucial, RSA encryption is used for securing the messages.

**Keywords —Bluetooth, Android Chat Application, Adhoc Network, MPR flooding, RSA.**

\*\*\*\*\*

## I. INTRODUCTION

In organizations or companies where communication is essential but due to prohibitions on the use of internet or due to unreachable cellular network; if one has to send a message to one of the colleagues or broadcast it to a group of them, then this can be done using a Bluetooth chat app. The objective of this app would be to send messages beyond the Bluetooth's normal range of 10m via the intermediary devices. The intermediary devices must also have the app running in the background. These devices can act as nodes buffering the message for certain time and then sending the message to their nearest neighbours till the message is reached to the destination. So as to achieve this, flooding algorithm is used. But using usual flooding have many limitations like a broadcast storm is created and then the complexity of the app increases for handling the duplicate messages. To avoid blind sending of messages, packets are sent to those nodes only which are most likely to fall in the destination route. This is achieved by Multi Point Relay (MPR) that is based on 2 hop neighbour knowledge of the node.

The main purpose of the app is to achieve network transparency and to allow secure communication between devices. For this type of communication, the message content would be encrypted with the public key that is exchanged when the devices are paired and will be decrypted at the destination with the private key that was generated along with public key. The public key and the private key are generated when the user logs in the app for the first time on a device.

The only limitation of this app arrives when the Bluetooth of the intermediary devices should be on, even though they are not sending messages and hence, the battery is getting consumed. However, as compared to other technologies like Wi-Fi, Bluetooth consumes lesser battery power and therefore, proves to be a better alternative for chat messenger.

## II. LITERATURE SURVEY

### A. Adhoc Network

A network that consists of independent nodes connected wirelessly such that those nodes have the ability form connections dynamically, such a network is called an Adhoc Network. A wireless ad hoc network (WANET) is a decentralized type of wireless network. In a Bluetooth network,



there are two types of nodes: a slave and a master. Each node has the ability to be either or both at the same time. Hence, we can say that, Wireless Ad-hoc Network is a set of wireless independent multi-hop nodes which does not require any pre-existing infrastructure. All wireless devices in the network, including the ones that are present outside the range of a particular node are discovered and peer-to-peer communication takes place between them using multi-hop technique. So, some of the nodes in ad-hoc network may not be able to communicate directly with each other and are dependent on some other nodes to pass their message. Such networks are often known as multi-hop or store and forward networks. The intermediate node(s) act as routers, which discover and maintain a table to forward the message to other nodes in the networks. Many significant applications of Ad-hoc networks include sharing internet without having an access point, exchanging files and data in a group, underwater sensor networks and disaster recovery.

As topology of a Wireless Adhoc network is not static, selection of routing process becomes difficult. Thus, routing algorithms like link state routing and distance vector routing are not efficient for Adhoc networking due to the reasons mentioned before. So, to overcome this problem, several ad-hoc routing protocols have been proposed[1][2][3][4][5]. These routing protocols can be classified as firstly, proactive algorithm that maintains routing tables that contains the lists of destinations and their corresponding routes where the tables are updated periodically by sharing of tables with the adjacent node. Secondly, Reactive Algorithm, also known as on-Demand routing where the sender will try to find the route to destination (using shortest path algorithm) only if it has to send some data. Lastly, Hybrid Routing Algorithm which is a combination of both reactive and proactive protocol where each node first maintains the routing tables (as in case of proactive protocol) and will also participate in on-demand routing by exchanging routing tables on request by sender node (reactive protocol).

### **B. Flooding**

This is the most common algorithm used for adhoc networks, as in Adhoc network every node acts like a server and a client, there is no dedicated server. Similarly, the flooding algorithm treats every node as a receiver and transmitter. It tries to forward the message to its neighbouring nodes except for the node it has received the message from. However the uncontrolled flooding will keep on routing packets indefinitely and therefore various techniques are used to control this flooding. One such technique is to broadcast the messages to only those nodes that are likely to be on the destination route called MPR flooding(Multi point Relay) discussed in the next section.

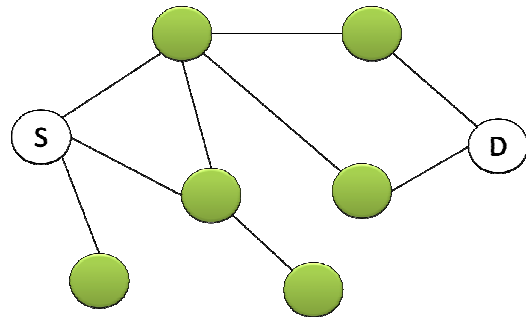


Fig. 1 nodes receiving packets from Blind Flooding.

### **C. Multipoint Relay flooding**

This algorithm [6] is based on 2 hop neighbour knowledge of the node implemented in the OLSR routing protocol[7]. Here the number of nodes retransmitting the messages are limited as compared to blind flooding. The nodes that are retransmitting these messages are called multipoint relay and then they decide their own MPR set to relay the messages. This keep on continuing till the destination is found.

The algorithm is as follows:

- (i) The sender will first make two sets of nodes, one being h1 (1 hop neighbours) and the second one being h2 (2 hop neighbours).
- (ii) Then it checks which h1 nodes have the h2 nodes as only their neighbours from the h1 node set and then adds that h1 node to MPR set.
- (iii) From the remaining h1 nodes it is then checked which one of them covers the

maximum of uncovered nodes of h2 and then adds that h1 node in MPRset. In case of, if two or more nodes from h1 set have same number of uncovered nodes of h2 then both covered and uncovered nodes are compared.

- (iv) This keep on continuing for every node in the MPR set.
- (v) The loop stops once the destination is reached.

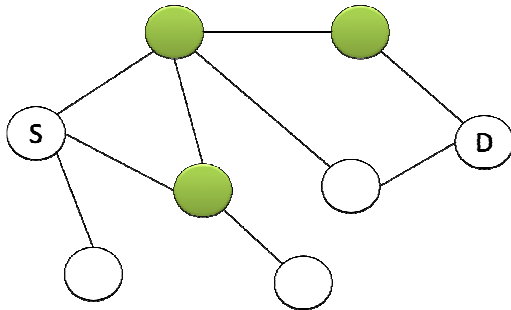


Fig. 2 nodes receiving packets from MPR Flooding.

#### D. Public Key Encryption

Encryption is the process of transforming a plain text data into encrypted form such that only the intended destination can decrypt or get the original data back from the encrypted data. Now, this process uses key (a piece of information or sequence of random elements that acts as a parameter being used in the algorithm). Keys are of 2 types, namely, public key and private (or secret) key. Public keys are the ones that are known to anybody who tries to communicate with the key-holder whereas the private keys are the ones that are only known to the key-holder. The public key and private key are mathematically related.

The main idea of encryption lies in the algorithm used to generate and use these keys and these algorithms are categorized into 2 categories: Asymmetric Key techniques and Symmetric key techniques. In symmetric, same key is used for encryption as well as decryption. As public key is available to both, the sender and the receiver, the public key is used for encryption and decryption. These types are comparatively less secure. However, in asymmetric, keys used

for encryption and decryption are not the same i.e. only the private key corresponding to a respective public key can decrypt the message encrypted by that public key. Asymmetric encryption is often referred to as Public key encryption.

Various public key encryption algorithms have been proposed. Some of the commonly used are Diffie–Hellman key exchange (used for secure key distribution), DSA (Digital signature algorithm) and RSA (for key generation, encryption and decryption). The algorithm that will be implemented here is RSA.

#### E. RSA

RSA is an algorithm that achieves both key generation as well as encryption and decryption. This algorithm is widely used for transmitting data securely.

The pseudo-code for the same can be given as follows:

- (i) Choose 2 prime numbers  $p$  &  $q$ .
- (ii) Compute  $n$  such that  $n = p * q$ .
- (iii) Compute  $\phi(n)$  as  $\phi(n) = (p - 1) * (q - 1)$ .
- (iv) Choose encryption factor 'e' such that  $1 < e < \phi(n)$  and  $e$  and  $n$  are co-prime.
- (v) Compute a value for decryption factor 'd' such that  $(d * e) \% \phi(n) = 1$ .
- (vi) Public Key is  $(e, n)$ .
- (vii) Private Key is  $(d, n)$ .
- (viii) The encryption of message 'm' is done using the public key which results in cipher text 'c'.
- (ix) The decryption of cipher text 'c' is done using the private key which results in message 'm'.

### III. PROPOSED SYSTEM

#### A. The Chat Application

1. At the installation of the app on the device, a public and a private key is generated using RSA algorithm.
2. Now, the device that wants to send the message to the destination device should be paired with the destination through the app

- at least once so as to retrieve the MAC address and the public key of the destination.
3. Before sending the message, the message is encrypted with the public key of the destination and the packet is structured. The Message format is described in the next section.
4. The message is then transmitted to the destination using MPR flooding algorithm, such that.
  - (i) If message reaches destination, an ACK is transmitted back to the sender.
  - (ii) If the destination was not found, an error report is generated.

#### B. Message Format

All the messages transmitted contain the sender Mac address and the receiver MAC address and the message content. The message has a time to live count attached to it i.e. the hop count, maximum hop counts it can travel before getting self-destructed. If the sender does not receive the ACK message from the destination for a certain period of time delivery error is reported. However this is possible when the destination phone could not be found or some hoe after receiving the message the destination device was turned off before it could send the ack.

Now, the messages exchanged need to be secured. For that, encryption techniques need to be used. The message format for the message that will be exchanged through the chat application is shown in figure 3.

Sender MAC Address	Receiver MAC Address	Message Content	Hop Count	ACK ( YES / NO )
--------------------------	----------------------------	--------------------	--------------	---------------------

Fig. 3 Message Format.

#### C. Encryption used for Chat Application

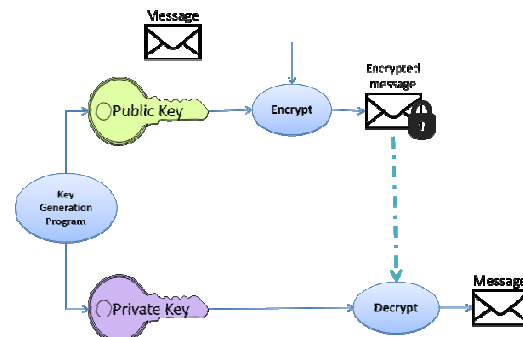


Fig. 3 Encryption used for Chat application.

Here, as soon as the app is installed (or user logs in) the app generates the public key and the private key for the user. For two devices to communicate, they must be paired. So, just after the pairing is done, the public key is shared. Now, looking from the destination's point of view, the public key will be used by the sender devices to encrypt the message that are to be sent to the destination device. When the destination device receives that message, the private key which was generated along with the public key will be used to decrypt it and obtain back the actual message at the destination end.

#### D. Limitations

Every device should have the app running. This app requires the Bluetooth to be enabled for infinite amount of time.

### IV. CONCLUSIONS

So, now with this chat application, people can communicate with each other via Bluetooth at places where internet and Wi-Fi chat applications cannot be afforded due to low internet connectivity or battery consumption problems. Being a chat application, security of messages is a critical issue here and hence the security provided needs to be enhanced.

### ACKNOWLEDGEMENTS

We wish to express our sincere gratitude to Dr. U. V. Bhosle, Principal and Prof. D. M. Dalgade, H.O.D of Information Technology Department of

RGIT for providing us an opportunity to do our project work on "Bluetooth Message Hopping". This project bears on imprint of many people. We sincerely thank our project guide Prof. Govind Wakure for his guidance and encouragement in successful completion of our project synopsis. We would also like to thank our staff members for their help in carrying out this project work. Finally, we would like to thank our colleagues and friends who helped us in completing the project synopsis successfully.

## REFERENCES

- [1] S. Basagni, I. Chlamtac, V.R. Syrotiuk and B.A. Woodward. A Distance Routing Effect Algorithm for Mobility (DREAM), Proceedings of the fourth annual mobile computing and networking, October 1998.
- [2] P. Krishna, M. Chatterjee, N.H. Vaidya and D.K. Pradhan. A Cluster-based Approach for Routing in Ad hoc Networks. In proceedings of Second USENIX Symposium on mobile and Location Independent Computing, pp. 1–10, January 1996.
- [3] S. Murthy and J.J. Garcia-Luna-Aceves. An Efficient Routing Protocol for Wire-Less Networks. ACM Mobile Networks and Applications, Special Issue on Routing in Mobile Communication Networks, 1(1):183–197, October 1996.
- [4] C.E. Perkins. Ad hoc on-demand distance vector routing, Internet Draft, Internet Engineering Task Force, work in progress, December 1997.
- [5] C.-H. Toh. A novel distributed routing protocol to support ad-hoc mobile computing, Proceeding of 15th IEEE Annual International Phoenix Conference on Computer Communications, pp. 480–486, 1996.
- [6] A. Qayyum, L. Viennot, and A. Laouiti. Multipoint relaying: An efficient technique for flooding in mobile wireless networks. 35th Annual Hawaii International Conference on System Sciences, 2001.
- [7] P. Jacquet, P. Muhlethaler, A. Qayyum, A. Laouitim, and L.Viennot. Optimized link state routing. draft-ietf-manet-olsr-06.txt, 2000.

# Urban Traffic Congestion Estimating Using Simplified CRONOS Model: Algorithm and Implementation

Abdallah Lakhouili<sup>1</sup>, Hicham Medromi<sup>2</sup>, El Hassan Essoufi<sup>1</sup>

<sup>1</sup> Faculty of Sciences and Techniques, Hassan First University, Casablanca road, Settat.

<sup>2</sup> High National School of Electricity and Mechanics, Hassan Second University, El Jadida road, Casablanca

\*\*\*\*\*

## Abstract:

In this paper, we present a method to calculate, the road urban traffic congestion using CRONOS model. For the model simulation we propose adapted algorithm. We compare our adaptive system results to others given by urban traffic system based on fixe pattern lights. The system is designed to make equilibrium between the length queue and the states of lights to search optimal strategy guarantying the fluidity of traffic and minimizing the total delay. The real-time urban traffic control model CRONOS has been applied on a simple intersection presenting two roads. We have choosing a simple intersection to validate our method which will apply after for a complicate intersection. We use traffic evaluation criteria to testing the accuracy the model. The results show benefits of the proposed model on the total delay compared to that of the fixe pattern lights system. All traffic situations, peak or low are concerned by these results.

**Keywords — CRONOS Model, Urban traffic congestion, intersections, traffic flow**

\*\*\*\*\*

## I. INTRODUCTION

The road traffic congestion of urban transport can be defined as the ratio of volume to the capacity of the road [19]. The volume and capacity (or demand) of the road cannot directly measured so the calculate value of congestion becomes subjective in nature. In this paper we propose a simple method based on a simplified CRONOS model and adapted algorithm to define optimal strategy to overcome the traffic congestion witch depends on the density and speed of the vehicles, are used for the estimation of road traffic congestion [18]. These two traffic parameters are considered by keeping in mind that general perception about the congestion on the roads increases when the number of traffic (traffic density) increases and also increases when the speed of the traffic decreases [20]. The density is the main input of our model and is defined as the number of vehicles entered in the intersection by second.

The output of the model is level of congestion represented by the length of waiting line or queue line. The road traffic congestion is one of the most confusing tasks, because there is no standard way of measuring congestion level on the roads and intersections [20]. It results in serious environmental, time wastage, health hazards, and economic problems. So that determining the congestion period and measuring traffic parameters which affects it are very important. In our study we use only one traffic parameters, the density, which affect mostly the congestion.

In section 2 we present a rapid overview of some approach used to estimate the urban traffic congestion and the main adaptive urban traffic control system. CRONOS model is presented in section 3. The proposal methodology to estimate the urban traffic congestion is given in section 4. In Section 5 we describe in detail the model simulation with variables and constraints of the urban traffic problem. The implementation of the model are described in section 6. The results and

### A. Different approach to estimate the congestion

discussion are presented in section 7. Finally, the conclusions of the present study are summarized in section 8.

## II. LITERATURE REVIEW

The modeling of traffic flows is an expensive and difficult task given the variability of flows and the amount of data to process. Several models have been developed to simulate and control the flow of traffic. We distinguish two main types of



models: macroscopic models and microscopic models. Macroscopic models describe globally the traffic flow, they are used in general in strategic study aims to giving planning or tactical solutions for the problems [8]. Although these models are fast, easy to calibrate and they do not require significant computer resources to simulate them, they are not well adapted to simulate realistic traffic flow, because the level of detail with which they deal the flow is very low. In contrast, microscopic models treat interactions between individual vehicles, they take into account many parameters related to vehicles, making the treatment of a large number of very delicate vehicles and calibrating a difficult task [21]. We cite as example: Krause and Alrock uses fuzzy logic to determine six discrete levels of congestion [13]. Porikli and Li determine five level of congestion from traffic flow information and video images using a Hidden Markov Model [17]. Jia and Li use different factors which affect the road traffic congestion [12]. Atikom and Pongpaibool estimates the road traffic congestion by using vehicle velocity [16].

### ***B. Adaptive urban traffic control system (UTC)***

To overcome the urban traffic congestion, urban traffic control system (UTC) using traffic signals is a major element of safety for vehicles and pedestrians when crossing an intersection. Traffic signals share the time between the different flows, that are considered to be antagonistic, and eliminate the most serious conflicts. This was the first reason for installing signals on intersections in the early 1900s [7]. Growing amount of traffic in urban areas made UTC by signals a significant traffic management tool in order to reduce the consequences of this increase: congestion, delay, stop, pollution, fuel consumption, noise, stress, discomfort. In the point view of methodology, the main problem of controlling signalized intersections is concedering different conflictual objectives. In fact; looking for the best fluidity is generally incompatible with the best safety in the sense that controlling the traffic from a safety point of view implies constraints on the traffic signal color durations and this implies limitations on traffic fluidity management. Competing interests in crossing an intersection of the different types of user (vehicle, pedestrian, emergency vehicle, bus, bicycle, etc.) made the task very difficult. Another problem of urban traffic management is to reach the optimum and search to carry out

the objectives for the best. A number of constraints appear:

flexibility in the control method, the size of the controlled network, the CPU time required, the reliability of sensors, modeling and forecasting, etc. Many UTC systems exist, all over the world, from the simplest fixed-time plan to the new generations which lead to greater flexibility of control. In our study we present the UTC model „control of networks by optimization of switchovers“ (CRONOS)[12],[5]. The authors of this model/algorithm have fixed as objectives on the one hand using advanced sensors for real-time traffic measurements and secondly focusing on its speed and flexibility.

The adaptive time plans are certainly the most widely used control systems in the world. When the plan is well adjusted to the recurrent traffic situations and the sensors are working correctly, these methods show good performances in a wide range of situations [7].

The first adaptive system that was applied is the English system SCOOT. This system produces a time plan incrementally. The Australian system SCATS builds a time plan piece by piece. These systems are very close to a classical time plan, but they lead to a new type of flexibility, which follows changes in traffic not only throughout the day but also over a period of months [15].

Recently other systems have appeared: CARS in Spain [2] and MOTION in Germany [4]. Other systems OPAC in the USA [10], PROLYN in France [11] and UTOPIA in Italy [9] had adopted a different strategy aims to minimize a traffic criterion using an optimization method to determine the green and red stage durations by time steps of 4 or 5 s. The cycle duration is not constrained and varies from one cycle to the next. These systems initially determine the different stages of the intersection and use minimum and maximum green durations. Thanks to their small time step and the use of magnetic-loop-based sensors, these systems take into account the traffic flow variations at a scale of a few seconds and more globally (at the level of the intersection) than the vehicle actuation functions [8].

This new generation of systems does not need to re-actualize the control system after a few years as in the case of time plans. Another advantage is their greater flexibility for finding the green and red durations according to the traffic situations, especially for those which have a wide possible cycle spectrum at each cycle.

Several experimental studies have shown benefits obtained by these UTC systems on the delay and the journey time compared to actualized time plans, although this result depends on the intersection characteristics and the traffic situations.

The CRONOS algorithm was developed in the 1990s to realize two objectives: build a non-exponential and fast

optimization method providing the traffic signal states for the next second in less than one second. The stake is to react as fast as possible to the traffic variations. This method will not look at all solutions but will use a heuristic providing a good local minimum. The second objective was to use image-processing-based traffic measurements, like queue length on link and vehicle spatial occupancy inside the intersection [1].

### III- THE CRONOS ALGORITHM

Consider a zone of several adjacent intersections.

The general process of the CRONOS system showed in figure 1 [7] can be summarized as follow: The one-second traffic measurements feed a forecasting and a modeling module. The forecasting module predicts, for a given time horizon, the future vehicle arrivals on each link entering the zone. This prediction is based on a rolling average of the arrivals in the past; it is used by the modeling module which calculates the value of a chosen traffic criterion for a given sequence of traffic signal states (colors) over the time horizon. These states are provided by an optimization module, which looks for the best sequence which minimizes the traffic criterion. When this sequence is found, the corresponding traffic signal states are applied on the intersection for the next time step, and the whole process is activated again one time step later [7].

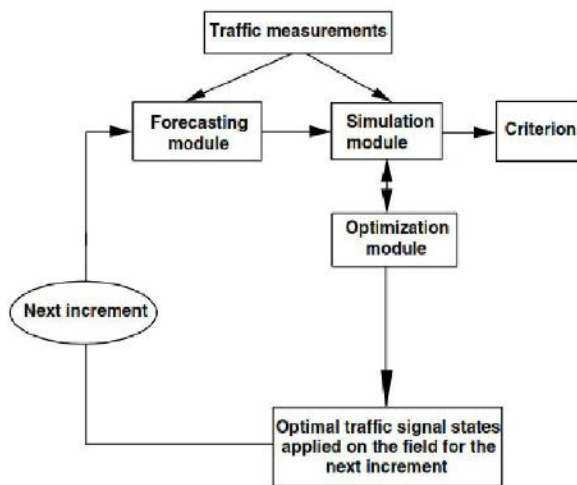


Figure 1 : General structure of CRONOS algorithm

The optimized traffic criterion is the total delay expressed on length of the waiting line the zone over the time horizon. The links entering the zone or connecting two intersections of the zone are considered in the delay. The criterion C is written as:

signal groups), traffic data (saturation flow, free flow) and traffic signal data (safety constraints).

$$\sum (\sum \sum )$$

Where  $q_{s,j}$  is the queue length at the time step  $s$  of the horizon  $H$  on the link  $j$ ,  $n_{s,k}$  is the number of stopped vehicles at the time step  $s$  in the storing inner area  $k$  of an intersection. The total number of links in the zone is  $N$ . The total number of storing areas inside the intersections is  $M$ .  $C$  is based on the queue evolution equation at each time step  $s$  according to the given arrivals vehicles and calculated departures during  $s$  ( $a_{s,j}$  and  $d_{s,j}$  respectively) for the considered link  $j$ :

A same equation is obtained for

The variables depend on the controlled variables which are the switchovers from red to green and green to amber (or red for pedestrian signals) of every traffic signal group. One group is defined as the set of traffic signals which controls the same traffic flow. At the initial time step  $s = 1$ , the queues and the storing areas inside the intersections are measured directly by the video sensors. The value of the rolling time horizon depends on the spatial extent of the zone. It is typically around one minute for one controlled intersection. The time-step value depends on the complexity of each controlled intersection and their number. It must be superior to the maximum computational time for solving the optimization process [7]. The optimization module of the CRONOS system calculate the optimized traffic criterion by a heuristic method based on a modified version of the Box algorithm [14]. The principle of the method is as follows: in the first step, the criterion value is calculated for a set of initial solutions (a solution is the set of values for the controlled variables over the time horizon). The second step is an iterative process up to convergence and consists, at each iteration, in looking for the worst solution and modifying it. Two types of modification are used: the first one tries to move the worst solution away from the centroid of the other solutions. The second one tries to bring the worst solution closer to the centroid. The effect of these successive iterations is to lead all solutions towards a region of the solution space. The convergence is reached when all solutions are very close to each other [7]. The traffic measurements are obtained by automatic image-processing of video cameras. These one-second measurements are the queue length at the stop line for each link, the traffic flow at the entries or the exits of the intersection, the spatial occupancy inside the intersection for pre-defined spatial areas. These areas represent storing zones behind a traffic signal or for left-turning vehicles. The data types concern infrastructure data (list of intersections, list of links, link length, list of traffic

### IV- METHODOLOGY

To validate the proposal method, we have choose a simple

intersection with two roads having one direction traffic as shown in figure 2.

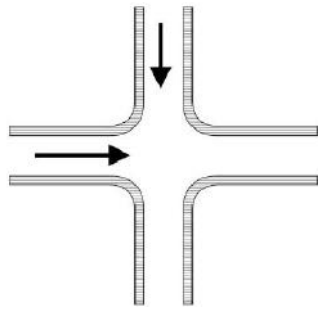


Figure 2 : Intersection design

Each link  $j$  of the intersection (there is one link for each road) is characterized by the input and output flow rate and at the time step  $s$  of the time horizon  $H$ . These values are given by simulators that we have developed.

The queue length at the time step  $s$  of the horizon  $H$  on the link  $j$  is and it is calculated by:

Where  $g$  is a boolean variable which is equal to 1 if the light is green and 0 if it's red and  $t_g$  is its duration.

For this configuration of intersection, the stopped vehicles in the storing inner area of intersection are not considered and is equal to 0. So the optimized traffic criterion is written:

$$\sum_j \left( \sum_s \right)$$

## V- TRAFFIC CONGESTION MODELING SIMULATION

To simulate the model we take into consideration some real traffic flow restrictions in the intersection and we define relative constants and variables.

### A. Agent-oriented programming languages and JADE environment

Based on agent-oriented programming language, JADE is a software platform that provides basic middleware-

### A. Variables

and variables are generally delivered by the magnetic sensor installed in the entry of links. In our case it is given by a programmed simulator. Variables depend on the controlled variables which are the switchovers from red to green and green to amber of every traffic signal group. We symbolized it by a boolean variable which is equal to 1 if the light is green and 0 if it's red and  $t_g$  is its duration.

At the initial time step  $s = 1$ , the queues and the storing areas inside the intersections are measured by the sensors. In our case they are initialized by the system. The value of the rolling time horizon generally depends on the spatial extent of the zone. We fixed it at one minute for one intersection. The time-step value depends on the complexity and number of intersections, it must be superior to the maximum computational time for solving the optimization process. The infrastructure is defined by list of intersections, list of links, link length, list of traffic signal groups.

### B. Constraints

The intersection is described as a set of safety constraints on the traffic signal groups. The main constraints are the minimum/maximum of the duration of each traffic signal state, antagonism criteria and the clearance time.

### C. Simulation algorithm

- For each horizon time  $H$ , constitute a number of matrix  $Q$  of dimension  $(6 \times 60)$
- Each 60 lines of matrix  $Q$ , correspond to vector  $R$   $(g_1, g_2, g_3, g_4, g_5, g_6)$  where  $i$ : input,  $o$ : output,  $s$ : time step,  $j, k$ : link of intersection.
- Calculate  $Q$  for each time step  $s$  of the horizon time  $H$  and each link  $j, k$ .
- or each iteration  $r$  calculate  $\sum_{s=1}^{60} (Q_{s,j,k})$  each iteration correspond to one vector  $R$
- Calculate  $\min (Q)$ .

## VI- MODEL IMPLEMENTATION.

To implement the proposed model we used the JADE framework environment, using JAVA/J2EE object-oriented programming language and framework.

layer functionalities which are independent of the specific application and which simplify the realization of distributed applications that exploit the software agent abstraction.[3],[22].

### B. Creation of Agents

The first phase of implementing the system involves the creation of some agents to meet the requirements of its corresponding link, intersection or task.

For the purposes of our case study to validate the system, the traffic network consists of one intersections. Each intersection

information provided by traffic sensors are the inputs to the others agents as indicated in agents description above.

In this implementation, the system administrator at runtime enters the initial conditions for each link of the intersection, providing a snap-shot of the environment. In a real-time environment, this would be modified by RegulaAgent wich will calculate a real state of urban traffic. This would allow the SectionAgent to update the stored values of . In addition to this information, RegulaAgent stores the current state of each traffic light. With the relatively small size of this traffic network implementation, we only use vectors to hold the data. Figure 3 and 4 show the Class diagram of agents and there interactions.

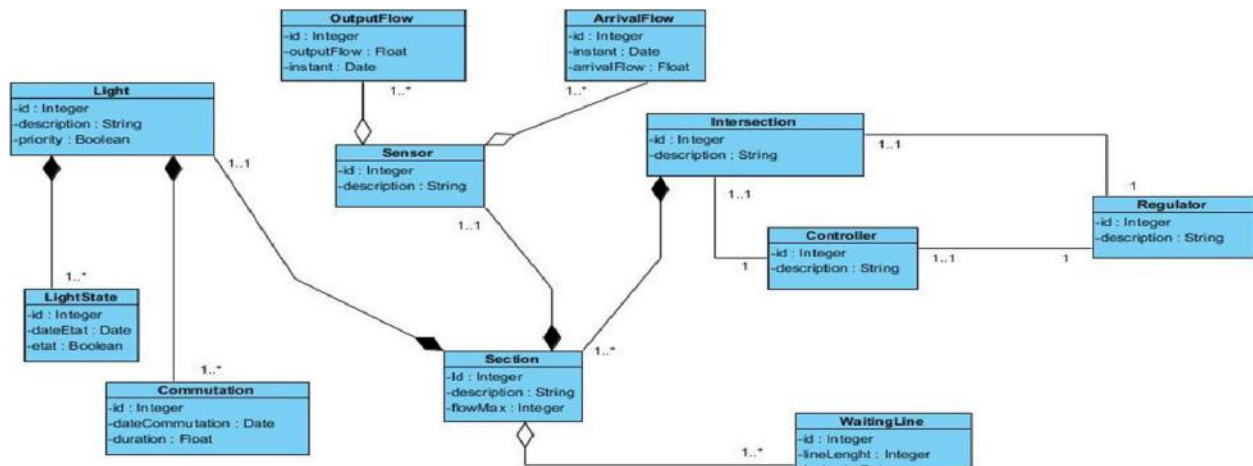


Figure 3 : Class diagram of agents

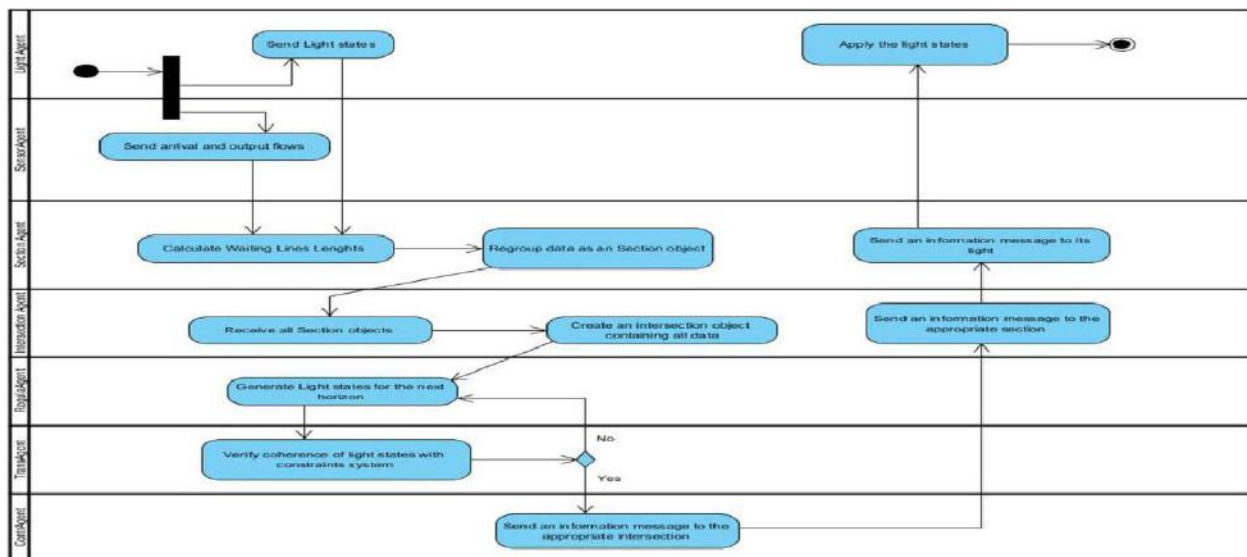


Figure 4: Agents interactions

calculated for each time horizon H on several sample of matrix to find optimal traffic criterion corresponding to low level of congestion. The obtained results, are compared to those given by the same system using fixe pattern light which. For this pattern light strategy we fixed green light duration to

To realize this comparison we calculate total delay obtained by summing every queue length, measured in terms of number of waiting vehicles at a traffic signal each second on the all link of intersection over the one entire hour.

Figures 5, 6, 7 show the principal results of simulation and the benefits of our system.

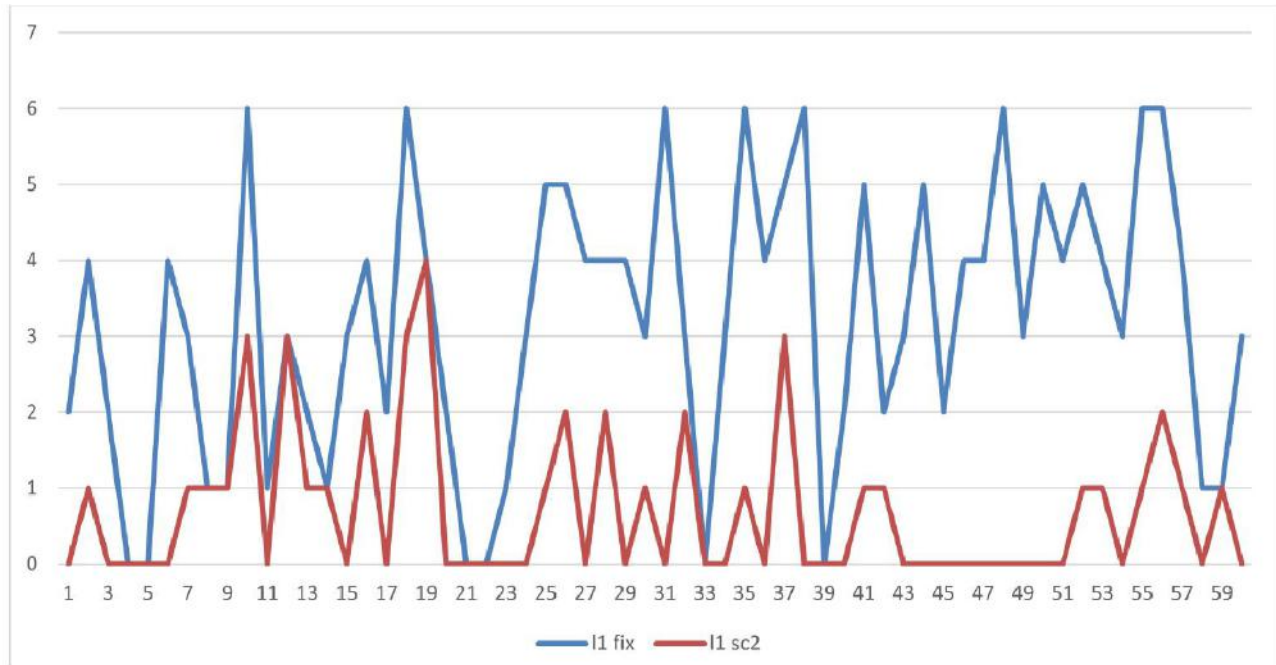


Figure 5: Link 1 length queue by fix pattern (blue) and proposed system (red)

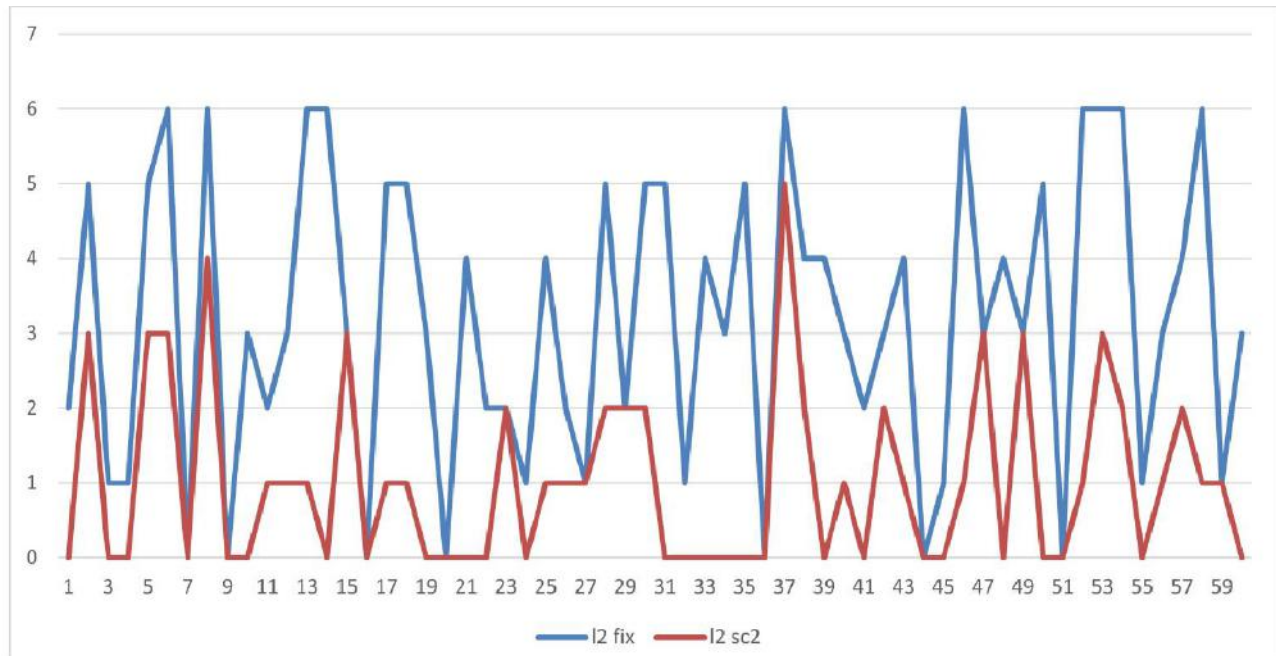


Figure 6: Link 2 length queue by fix pattern (blue) and proposed system (red)



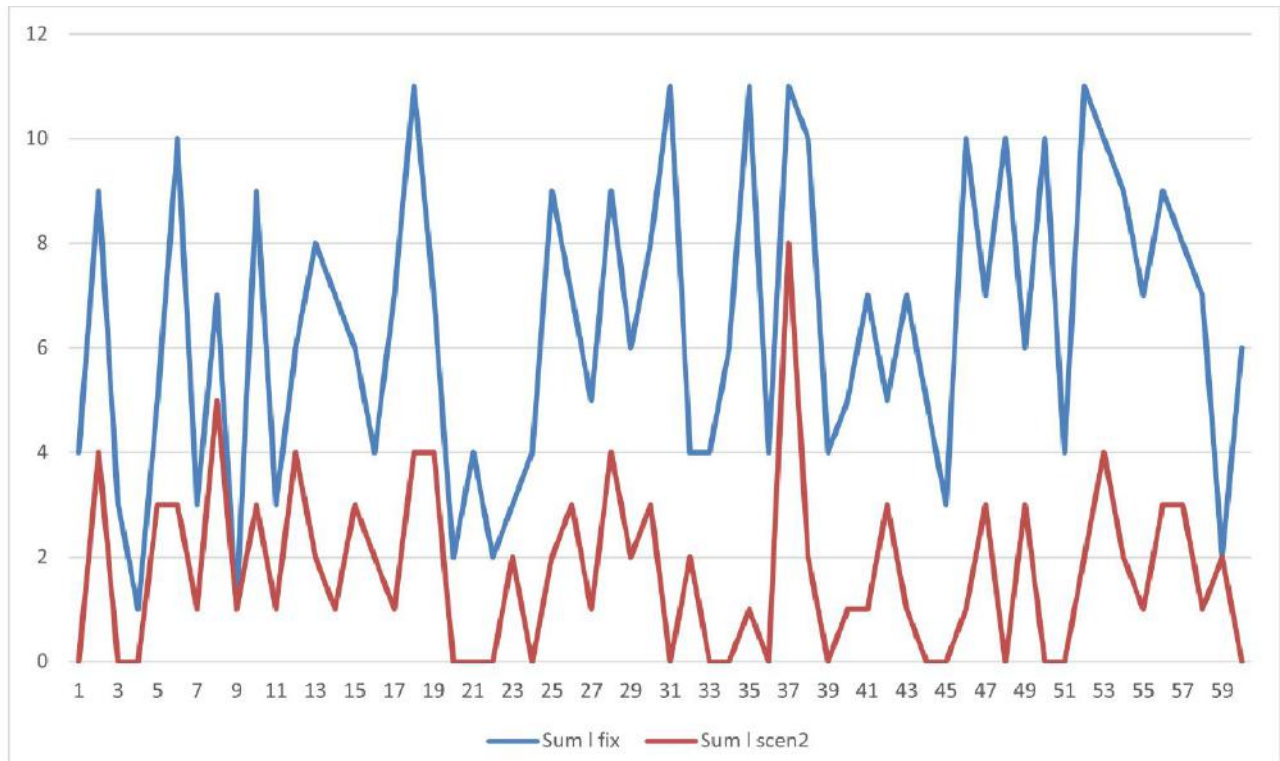


Figure 7: Sum of length queue by fix pattern (blue) and proposed system (red)

## VIII. CONCLUSION

In this study, we proposed a method to calculate and optimize the urban traffic congestion in the intersection. We adopt a simple case to validate the system and methodology. In a future work we will developing our system and intersection sample to take in consideration a real situation of urban traffic and taken into account a maximum of urban traffic variables to improve system accuracy. Generally The real-time CRONOS strategy gives higher benefits on almost all of the traffic variables studied whatever the traffic situation peak or low. These opportunity is due to the CRONOS characteristics, especially its high flexibility and its global traffic optimization of the intersection, and due to the use of real-time video-based measurements witch are simulated in our case.

## REFERENCES

- [1] Aubert, D., Bouzar, S., Lenoir, F., Blosseville, J.M. *Automatic vehicle queue measurement at intersections using image-processing*. In: 8th International Conference on Road Traffic Monitoring & Control, London, vol. 422, pp. 100–104. 1996
- [2] Barcelo, J., Grau, R., Eagea, P., Benedito, S., 1991. *CARS: A demand responsive traffic control system*. In: Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, pp. 91–95.
- [3] Bellifemine, F., Caire, G., Greenwood, D. *Developing Multi-Agent Systems with JADE*. Wiley Series in Agent Technology. Michael Wooldridge, Liverpool University, UK. 2007.
- [4] Bielefeldt, C., Busch, F. *MOTION: a new one-line traffic signal network control system*. *Road Traffic Monitoring and Control* IEE. London 391, 55–59. 1994.
- [5] Boillot, F., Blosseville, J.M., Lesort, J.B., Motyka, V., Papageorgiou, M., Sellam, S., 1992. *Optimal signal control of urban traffic networks*. In: 6th International Conference on Road Traffic Monitoring and Control, IEE, London, vol. 355, pp. 75–79.
- [6] Boillot, F., Braban, C. *Les systèmes temps réel de commande de feu en milieu urbain*. INREST. Synthèse N° 44. Avril 2003.
- [7] Boillot, F., Midenet, S., Pierrelée, J-C *The real-time urban traffic control system CRONOS: Algorithm and experiments*. *Transportation Research Part C* 14 (2006) 18–38.
- [8] BUISSON C. LEBACQUE J. P. LESORT J. B. *STRADA, a discretized macroscopic model of vehicular traffic flow in complex networks based on the godunov scheme*. *Computational Engineering in Systems Applications*, pages 976–981, 9-12. July 1996.
- [9] Donati, F., Mauro, V., Roncolini, G., Vallauri, M., 1984. *A hierarchical decentralized traffic light control system*. The First Realization: Progetto Torino. In: 9th World Congress of the International Federation of Automatic Control, Budapest, vol II 11G/A-1.
- [10] Gartner, N.H.,. *Development and testing of a demand responsive strategy for traffic signal control*. In: Proceedings of American Control Conference, pp. 578–583. 1982.
- [11] Henry, J.J., Farges, F., Tuffal, J. *The PROLYN real time traffic algorithm*. In: 4th IFAC, IFIP, IFORS Conference on Control in Transportation Systems, Baden, pp. 307–312. 1983.
- [12] Jia, L., Li, C., “*Congestion Evaluation from Traffic Flow Information Based on Fuzzy Logic*,” *IEEE Intelligent Transportation Systems*, Vol. 1, 2003, pp. 50-53.
- [13] Krause, B., Altrock, C. “*Intelligent Highway by Fuzzy Logic: Congestion Detection and Traffic Control on Multi-Lane Roads with Variable Road Signs*,” 5th International Conference on Fuzzy Systems, New Orleans, 8-11 September 1996, pp. 1832-1837.
- [14] Kuester, J.L., Mize, J.H.. *Optimization Techniques with Fortran*. McGraw-Hill Book Company, pp. 368–385. 1973
- [15] Lowrie, P.R. *The Sidney co-ordinated adaptive traffic system: principles, methodology, and algorithms*. In: Proceedings of the IEE Conference on Road Traffic Signalling, London, vol. 207, pp. 67–70. 1982.
- [16] Pattara-Atikom, W., Pongpaibool, P. “*Estimating Road Traffic Congestion Using Vehicle Velocity*,” *Proceeding of 6th International Conference on Telecommunications*, Chengdu, June 2006, pp. 1001-1004.
- [17] Porikli, F., Li, X., “*Traffic Congestion Estimation Using HMM Models without Vehicle Tracking*” *IEEE Intelligent Vehicles Symposium*, Parma, 14-17 June 2004, pp. 188-193.
- [18] Posawang, P., Phosaard, S. *Perception-based Road Traffic Congestion Classification using Neural Networks*. Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, 2009, London, U.K.
- [19] Public Roads office, “*Highway Capacity Manual: Practical applications of Research*,” US Department of Commerce, Washington DC 1950.
- [20] Shankar, H., Raju, P., Rao, P. *Multi Model Criteria for the Estimation of Road Traffic Congestion from Traffic Flow Information Based on Fuzzy Logic*. *Journal of Transportation Technologies*, 2012, 2, 50-62.
- [21] Wagner, P., Brockfeld, F. *Testing and benchmarking of microscopic traffic flow simulation models*. *WCTRConference*, July 4th-7th 2004.
- [22] Wooldridge, M.J., Jennings, N.R. *Intelligent Agents: Theory and Practice*. In *Knowledge Engineering Review*. 1995