International Journal of Computer Science & Emerging Technologies

ISSN 2044-6004 (Online)

Volume 2 Issue 1

February, 2011

www.ijcset.excelingtech.co.uk

ijcset@gmail.com

IJCSET BOARD MEMBERS

- **Nafiz Imtiaz Bin Hamid**, Bangladesh
- **Jasvir Singh**, India
- **Manas Ranjan Biswal**, India
- **Ratnadeep R. Deshmukh**, India
- **Sujni Paul**, India
- **Mousa Demba**, Saudi Arabia
- **Yasser Alginahi**, Saudi Arabia
- **Tarun Kumar**, India
- **Alessandro Agostini**, Saudi Arabia

TABLE OF CONTENTS

1. **SEMANTIC QUERY EXPANSION USING KNOWLEDGE BASED FOR IMAGES SEARCH AND RETRIEVAL 1**

   Roohullah Khan, J. Jaafar

2. **TWO-LAYERED FUZZY LOGIC BASED FREQUENCY MODULATION CONTROLLER FOR A TWO-AREA THERMAL REHEAT POWER SYSTEM INTERCONNECTED WITH AN AC/DC HYBRID TRANSMISSION SYSTEM 6**

   V.Adhimoorthy, Dr.I.A.Chidambaram

3. **EFFECT OF RADIO PLATFORM NOISE ON EXISTING MAC PROTOCOLS IN USE IN AD HOC WIRELESS NETWORKS AND ITS SOLUTION 13**

   Prabir Banerjee, D.K.Basu, M.Nasipuri

4. **VARIATIONAL ITERATION METHOD FOR SOLVING NONLINEAR FRACTIONAL INTEGRO-DIFFERENTIAL EQUATIONS 18**

   Muhammet Kurulay, Aydin SECER

5. **PHISHING E-MAIL ANALYSIS 21**

   Shamal Firake, Pravin Soni, Dr. B.B.Meshram

6. **A STUDY ON SEARCHING MECHANISMS IN SEMANTIC WEB 34**

   M.Thangaraj, G.Sujatha

7. **BRIDGING THE GAP BETWEEN UML AND HARDWARE DESCRIPTION LANGUAGES AT EARLY STAGES OF EMBEDDED SYSTEMS DEVELOPMENT. 45**

   Fateh Boutekkouk

8. **AN ENERGY-EFFICIENT UNICAST ROUTING PROTOCOL FOR WIRELESS SENSOR NETWORKS.......... 60**

   Young-Jun Chung

9. **THE IMPACT OF THE DIGITAL DIVIDE ON EDUCATION AND HEALTH......... 65**

   Veena Paliwal, Vaibhav Sharma

10. **A NEW NARROW-BLOCK MODE OF OPERATION FOR DISK ENCOMPRESSION WITH TWEAKED BLOCK CHAINING 71**

Debasis Gountia, Dipanwita Roy Chowdhury

11. IMPACT OF WORMHOLE ATTACKS ON MANETS 77

Saurabh Upadhyay, Brijesh Kumar Chaurasia

12. CONTENT ANALYSIS OF MATHEMATICS WEBSITES IN TAIWAN 83

Hsiu-fei Lee

13. ANTICIPATING FEATURES OF RING FINGER FROM MIDDLE FINGER WIDTH: A NOVEL METHOD 89

Manimala.S, C.N.Ravi Kumar

14. INFORMATION EXTRACTION FROM REMOTE SENSING IMAGE (RSI) FOR A COASTAL ENVIRONMENT ALONG A SELECTED COASTLINE OF TAMILNADU 95

K. Bhuvaneswari, R. Dhamotharan, N. Radhakrishnan

15. DESIGNING FLEXIBLE GUI TO INCREASE THE ACCEPTANCE RATE OF PRODUCT DATA MANAGEMENT SYSTEMS IN INDUSTRY 100

Zeeshan Ahmed

16. PROCESS ORCHESTRATION FOR INTRUSION DETECTION SYSTEM BASED ON SOA AND EVENT DRIVEN ARCHITECTURE PRINCIPLES 110

K.V.S.N.Rama Rao, Pandu Prudhvi, Manas Ranjan Patra

17. PERFORMANCE ANALYSIS OF PARALLEL MINING FOR ASSOCIATION RULES ON HETEROGENEOUS SYSTEM 124

Rakhi Garg, P.K.Mishra

18. COMPARATIVE GENOME ANALYSIS USING COMPUTATIONAL APPROACH IN HIV-1 AND HIV-2 129

Dr. DSVGK Kaladhar, A.Krishna Chaitanya

19. A STUDY OF NEUTROSOPHIC TECHNOLOGY TO RETRIEVE QUERIES IN RELATIONAL DATABASE 133

Ashit Kumar Dutta, Ranjit Biswas, Nasser Saad AL-Arifi

20. AUTOMATIC DECIPHERMENT OF ANCIENT INDIAN EPIGRAPHICAL SCRIPTS - A BRIEF REVIEW 139

Soumya A, G Hemantha Kumar

21. AN ANALYSIS OF SOFTWARE COST ESTIMATION TRADITIONAL MODELS WITH NEURAL NETWORK BASED APPROACH 144

Manpreet Kaur, Sushil Garg, Mamta Sharma

22. CONTENT BASED RECOMMENDER SYSTEMS 148

P. Deivendran, Dr. T. Mala, B.Shanmugasundaram

23. GENERIC TRUST-BASED RESOURCE BROKER ARCHITECTURE FOR GRID 153

Damandeep Kaur, Dr Jyotsna Sengupta

24.EFFECTIVE CLASSIFICATION TECHNIQUE ENHANCED USING GENETIC ALGORITHM:FOR DATA MINING DISEASE IN THE INCUMBENTS TO THE HEALTH CENTRE 157

Manaswini Pradhan, Dr. Ranjit Kumar Sahu

25. SPEECH ENHANCEMENT USING GRADIENT BASED VARIABLE STEP SIZE ADAPTIVE FILTERING TECHNIQUES 168

G.V.S.Karthik , M. Ajay Kumar, Md. Zia Ur Rahman

26. HIGH PERFORMANCE HUMAN FACE RECOGNITION USING INDEPENDENT HIGH INTENSITY GABOR WAVELET RESPONSES: A STATISTICAL APPROACH 178

Arindam Kar, Debotosh Bhattacharjee, Dipak Kumar Basu, Mita Nasipuri, Mahantapas Kundu

27. A NEW PROTOCOL FOR ADVANCED ENCRYPTION/DECRYPTION USING ECC & PHAL 188

Yogendra Kumar Jain, Nitin

28. HIGH DIMENSIONAL DATA CLASSIFICATION THROUGH ATTRIBUTE REDUCTION USING RPCA APPROACH 196

Rajashree Dash, Rasmita Dash

29. STEGANOGRAPHIC TOOLS FOR BMP IMAGE FORMAT 200

Prof.Sumedha Sirsikar, Prof. Asavari Deshpande

Semantic Query Expansion Using Knowledge Based for Images Search and Retrieval

Roohullah Khan[1] and J. Jaafar[2]

Department of Computer and Information Sciences
UniversitiTeknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh,Perak, Malaysia.

[1]Roohullah_orc@yahoo.com, [2]jafreez@petronas.edu.my

Abstract—*The falling prices of multimedia and storage devices make almost everyone to act like a professional to capture photo and archive them for later use. Without efficient retrieval methods the search of images in large collections can become a painstaking work. Most of the traditional image search engines rely on keyword-based annotations which lacks the query semantic space equivalent to the annotation semantic space, because of the difficulty in describing the same concepts with other keywords. In this paper, we propose a novel approach for the query expansion using lexical and commonsensical knowledgebases like WordNet and ConceptNet, which will not only fill the gap in the semantic space between user query and annotation but will also provide an opportunity to discard the less important words from the query semantic space. For evaluation we have selected LabelMe datasets, which is openly available for researcher.*

Keywords**:** *Content Based Retrieval, Semantic Gap, Query Expansion, WordNet, ConceptNet.*

1. Introduction

With the increase of the digital media both online and offline, there is a growing increase in the demand for the system that can process, store, organise and manage the digital media efficiently and effectively. Processing and managing such an ever increasing amount of data is a great challenge. Keeping this, it is impossible for the user to manually search the relevant images from the large image corpus. This explosive growth of the digital media [1] without appropriate management mimics its use.

Currently, the multimedia search and retrieval are an active research dilemma among the academia and the industry. The large data is available an online repositories like Google, YouTube, Flicker, etc. provides required images or videos through text based matching techniques [31] (surrounded text around the images likes an image name, metadata) in which the novice user has hardly found and accessing the useful images or videos of interest and becomes difficult. Finding the image of interest becomes harder and harder. The area of the textual information retrieval is matured. Nevertheless, the image retrieval is still worth investigating. Due to this explosive growth, there is a strong urge for the system that can efficiently and effectively interpret the user demand for searching and retrieving the relevant information images.

Based on above reasons new semantic Query Expansion techniques using knowledge based for Images search and retrieval will be proposed and develop. This technique should be able to convert a user demand into set of discrete concepts. The semantic query algorithm which would be automatically interpreting the query according to the user's requirements.

The proposed technique will expand the query and interprets the user query semantically so that it can be further processed for the accurate retrieval. The proposed query engine is the text based query engine that will expand the user query by combining WordNet [19], ConceptNet [20] knowledge bases to retrieve the results semantically with higher accuracy.

The image is ultimately a congregation of objects depicts some concepts. For a computer, an image is just affection a mix of pixels that are characterized by the low-level features like colour, shape, texture, etc. while for the human it is more than that. For human an image is the mix of one or more semantic idea. For them, it refers to, not the content of the image that's appearing, nevertheless rather a semantic idea that it representing. It is worth saying that for the same images dissimilar mankind extractsseveralconcepts depend on the nature and knowledge of the human. Due to the open-ended nature of the human and the hard coded computer nature there appears a problem known as the semantic gap. Which are showing in Figure 1. Semantic gap is one of the key problems between the Computer interpretation and human understanding for visual information.

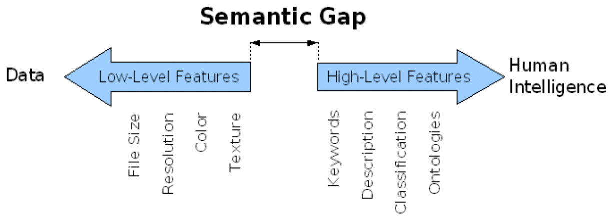


Figure 1. Semantic Gap

According to Smeulders et. al. [32] Semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation".

Although the start-of-the-art image retrieval techniques or the Content Based Images Retrieval (CBIR) system are a day by the day getting more and more powerful and can now achieve the accuracy up to some extent. Extensive research effort was done in retrieving the image assuming their visual content such as Query By Image Content (QBIC) [24], Netra [25], VisualSeek [26], WeebSeek [27], Virage[28], Videoq[29], Multimedia Analysis and Retrieval System (MARS) [30]etc. Indeed, the power of these tools doesn't reduce the semantic gap. And now the trend is completely moved from low-level features to the high-level idea.

The process of extracting the semantic idea from the images through hard coded machine remains a difficult it will be creating the following general research questions.

i. How can we interpret the user requirements that are given to the system in the form of query?

ii. How to reduce the semantic gap for Images of human intelligent and machine interpretation?

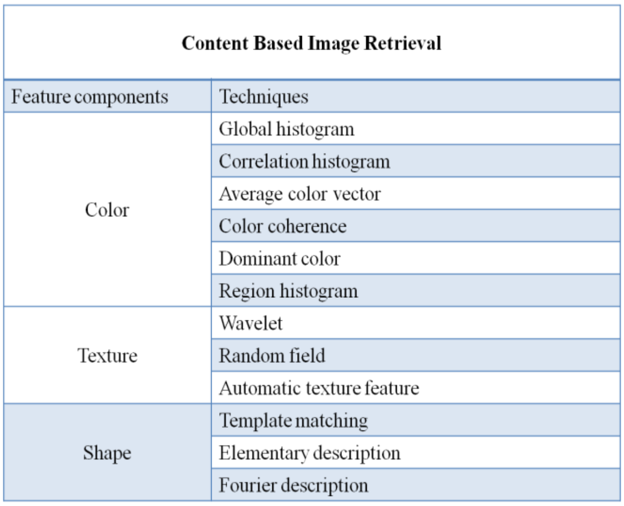
iii. How to bring the correctness to the system result along with the semantic proficiency?

2. Related Work

Content Based Retrieval (CBR) system is the widely uses for the Image retrieval but this system has retrieved the images through Content based query. CBIR is the techniques that retrieve the images through low-level feature like colour, shape and texture.

CBIR uses the visual content of image such as color, shape, texture, etc. some fundamental techniques for content based image retrieval include visual content description, similarity distance measures, indexing scheme as shown in the Table 1. Some retrieval system has incorporated user relevance feedback in order to facilitate the retrieval process.

Table 1: Depicts the content based image retrieval by using low-level features



Content Based Image Retrieval	
Feature components	Techniques
Color	Global histogram
	Correlation histogram
	Average color vector
	Color coherence
	Dominant color
	Region histogram
Texture	Wavelet
	Random field
	Automatic texture feature
Shape	Template matching
	Elementary description
	Fourier description

Query by Colour is the most prominent visual feature in CBIR since it is well correlated with human visual perceptions of objects in an image. Retrieving images based on color [7], [8] similarity is achieved by the color histogram for each image. A color histogram is a representation of the distribution of colors in an image, derived by counting the number of pixels of each of a given set of color ranges in a typically two-dimensional (2D) or three-dimensional (3D) color space. Color searches will usually involve comparing color histograms. When comparing is matching then retrieve that type of images. Similarly query by shape CBIR is also matching the Retrieving the images based on the shape include the shape [9], [10] of a particular region that is being sought out. While for the Texture CBIR look for visual patterns in images[11], [12] and how they are spatially defined. Textures are represented by texels or texture elements (also texture pixel), and retrieve the images through texture, wavelets. That the same texture matches to other images. However, the Deok-Hwan Kim andSeung-Hoon Yu [13] have been using CBIR through the region based instead to match the entire image. However, they can also get low level precision from the entire results. However,Ying Liu et al [7] have also been using the region based content matching and get little more accurate results. Nevertheless, the region based query has retrieved the images in a specific part of a region exist so through this query a lot of irrelative images have been retrieved, and the accuracy will be lower.

Query by example [10] is matching the input image as a hole through colour, shape, and texture instead as a part, and retrieve that images which are totally matching to this image. A query – by-example (query-by-image) retrieves the images based on the low level features. However, the visual feature lacks the semantics. Sometimes the images are visually similar, but it cannot contain similar information as shown in the Figure 2. While sometimes the images may be visually different but contain the same information as shown in the Figure 3. All these led to the semantic gap.

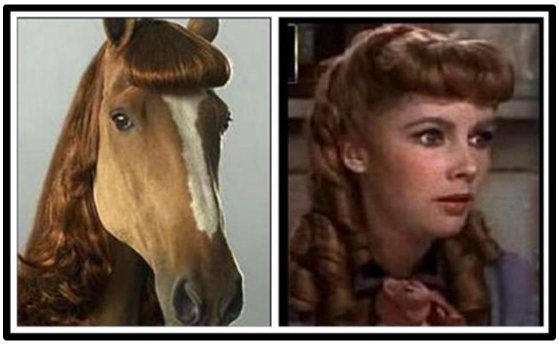


Figure 2. Depicts the visually similar but semantically different images

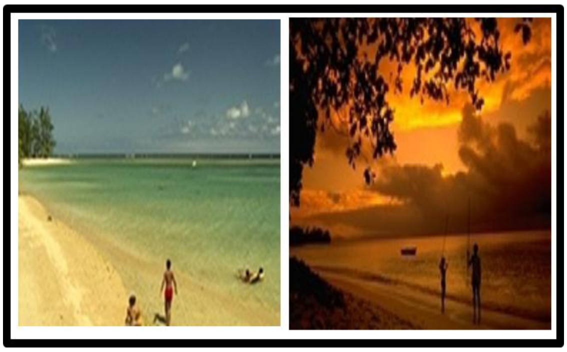


Figure 3. Depicts the visually different but semantically similar images

Furthermore, others query techniques have been using to find images from the large image corpus like querying by visual sketch [14], querying by direct specification for image features [15] and multimodal queries [13] (e.g. that combining touch, voice, etc.). However, all these CBIR techniques have not retrieved the images through semantic similarity? Now the trend has completely changed from the low level feature toward a high level semantic idea.

Semantics defines the concepts at a high level such as the objects, events, scene and the relationship among them.E.g. “Burning of wood in the street”. Where the wood is an object burning is an event, and street is a scene. It tells us that “What is actually happening in the image”.

Query plays a vital role in the performance of the information retrieval systems. Sometimes the user queries cannot define about they actually needs or sometimes the vocabulary in the query is inconsistent [6] with that in the relevant document. In order to solve this problem the general purpose knowledge bases are used such as WordNet, ConceptNet. Query expansion [2]-[5] for text based searching is too much successfully and given more accurate results with higher accuracy. This idea of the WordNet with query expansion is firstly, implemented by Zhiguo Gong et al [16], where WordNetis used as the basic expansion rules and then usesWordNet Lexical Chains and semantic similarity to assign terms in the same query into different groups with respect to their semantic similarities. Yokoyama et al [17], expand the query for the new users surfing the internet, they expand the query terms by WordNet, and the expanded query is then submitting to the search engine for getting most related web results. Ming-Hung Hsu et al. [18] combine the WordNet and ConceptNet knowledge bases for query term expansion.

WordNet is a lexical database for the English language [19]. It has developed by the George A. Miller under supervision. WordNetgroup's English words into a set of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The latest version ofWordNethasto contain over 155.00 words, grouped into over 117.00 synsets.

ConceptNet is a freely available, machine-usable common sense resource [20]. ConceptNet is developed by MIT Media Laboratory and is presently the largest commonsense knowledgebase. It is the relational semantic network that is automatically generated from about 700,000 English sentences of the Open Mind Common Sense (OMCS) corpus ConceptNet 3 presently consist of over 250,000 elements of common sense knowledge. The ConceptNet aims to give computer access to common-sense knowledge, the kind of information that ordinary people know [21].

3. Propose Method

Taking into account all the above issues we proposed a query expansion technique that will expand the user query lexically as well as conceptually to reduce the semantic gap. The user queries can be expanded lexically by using an open source lexical knowledge base, i.e. WordNet while the conceptual expansion can be done through ConceptNet. The overall framework of the proposed model is shown in Figure 4.

Throughout this model the user will be getting the images through searching from the large image corpus semantically. So this model is suitable to minimize the semantic gap between human understanding and machine interpretation. This model is also help to the user whoselects those words from the expanded query those are more relatives to the user original query and implementation for the searching. While stop those words which are fewer degrees of relevance have related to the original query.

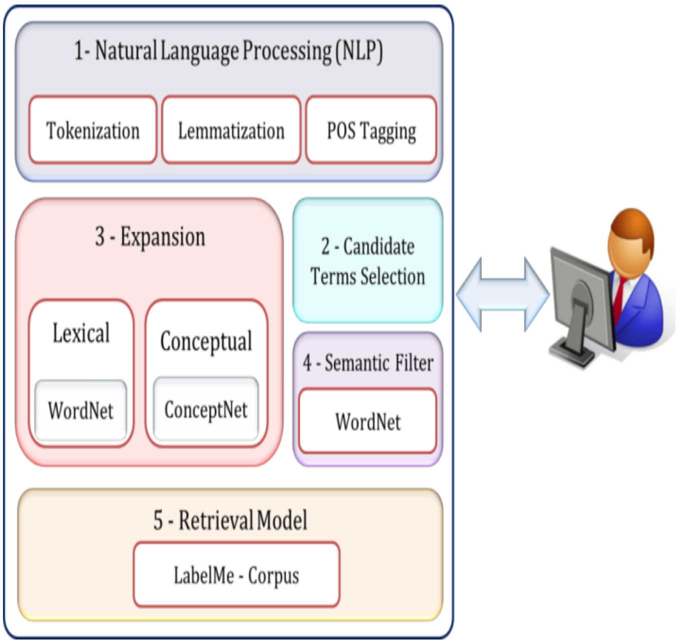


Figure 4. Depicts the overall model of the proposed framework

In this model, there are five (5) parts that get the user query and passing through this each step and after passing these steps the result will be display to a user. These five parts have to discuss details as below:

1. *Nature Language Processing:* Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. There are further three parts of this part.
    - *Tokenization:* To break the sentence into words is called tokenization that each word separate from the sentence.
    - *Lemmatization:* In linguistics, lemmatization is the process of grouping together the different inflected forms of a word, so they can be analysed as a single item or based form.
    - *POSTage:* is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context — i.e. Relationship with adjacent and related words in a phrase, sentence, or paragraph?
2. *Candidate term selection:* Candidate term selection means that just select the noun, verbs and adverb from the user search query for expanding.
3. *Expansion:* To expand the query through lexically from WordNet and Conceptually from ConceptNetto add more relative terms to the user original query.
4. *Semantic filter:* Semantic filters mean that filter the words matching with query with a select degree of relevancies that's stopping the irrelative or unusual words after the expansion and get the semantic relative words from the expanded.
5. *Retrieval Model:* In retrieving a model the search engine will be retrieving the result from the large image corpus that stores the images with relative annotations.

This model will be implemented in the label dataset [22]. Which are containing 31.8 GB datasets that containa total of 181, 932 images with 56946 annotated images, 352475 annotated objects and total of 12126 classes? The result will be containing through Vector Space Model.

Vector space model [23] is the information retrieval for the algorithm and framework.This model is an algebraic model to representing documents as a vector of identifiers. Vector space model is used in information filtering, information retrieval, indexing and relevancy rankings.The terms of a query surrogate can be weighted to take into account their importance, and they are computed by using the statistical distributions of the terms in the collection and in the documents [Salton 1983]. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document.

This model represents documents and queries as vectors.

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, ..., w_{t,q})$$

Each dimension corresponds to a separate term [33]. If a term occurs in the document, its value in the vector is non-zero.

Relevance rankings of documents in a keyword search can be calculated, using the assumptions of document similarities, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as a same kind of vector as the documents.

To calculate the relevance ranking of document Cosine is very easily calculating the angle between the vectors instead of the angle [34].

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

If the cosine value is equal to zero means that the query and document vector is orthogonal and have no match, i.e. no documents have match to the query.

The vector space model makes the following assumptions:

i. The more similar a document vector is to a query vector, the more likely it is that the document is relevant to that query.
ii. The words used to define the dimensions of the space are orthogonal or independent. While it is a reasonable first approximation, the assumption that words are pairwise independent is not realistic.

Vector space model is same to Boolean model, but it's the advance model. Which are some advantages?

i. Simple model to represent results based on linear algebra.
ii. Term weights not like a Binary model.
iii. Allow computing a continuous degree of similarity between query terms and documents.
iv. Allow ranking documents according to their possible relevance.
v. Allow partial matching of query terms to the documents.

4. Discussion

The main limitation of the past approaches was that the content based retrieval is an automatic solution for theretrieval, but they rely only on the low level feature's extraction. Now the question is that whether the low level feature extraction alone is enough for efficient searching and retrieval?

The answers might be no, because the low level feature only captures one aspect of the multimedia data. In addition, sometimes the images or videos that look similar are not semantically similar. So the retrieval results that are solely based on low level feature extraction are mostly unsatisfactory and unpredictable. This opens a new era for the research community to diverge from the existing methodologies to new a paradigm or new direction that there is something behind the visual features that need to be considered for accurate searching and retrieval.

That is the semantic of the multimedia data, i.e. high level features. Modeling the high level features are difficult than the low level features as the low level features are totally based on the colour, shape, texture structure while the high level feature isdepending on the semantics.

5. Conclusion

In a nutshell, they propose model use query expansion techniques for extracting the semantics from the user query. The model an effective way interpreting the user demand keeping in view the flexible nature of human as well as hard coded nature of computer. This model will be reducing the semantic gap by interpreting the user demand semantically in order to achieve the semantic accuracy as well as the efficiency in the retrieval.

Reference

[1] IDC white paper, "The diverse and exploding Digital Universe", sponsored by EMC. March 2008.

[2] J.Li, M. Guo and S. Tian "A New Approach To Query Expansion" 4th international conf. on Machine learning and cybernetics, Guangzhou, IEEE, 2005.

[3] Z. Yu, Z. Zheng, S. Tang and J. Guo "Query Expansion for answer document retrieval in chinese question answering system" 4th international conf. and cybernetics, Guangzhou, IEEE, 2005.

[4] H. Cui, J. Wen, J. Nie and W. Ma "Query Expansion by Mining User Logs" IEEE, vol. 15, No 4, pp. 829-839, 2003.

[5] B. Sun, P. Liu and Y. Zheng "Short Query Refinement with Query Derivation" Springer-Verlag Berlin Heidelberg, pp. 620-625, 2008

[6] JinxiXu., "Solving the word mismatch problem through automatic text analysis.", Ph.D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, MA, USA, May 1997.

[7] Y. Liu, X. Chen, C. Zhang and A. Sprague "Semantic clustering for region-based image retrieval" J. Vis. Commun. Image R. 20, pp. 157–166, 2009.

[8] M. ElAlami "Supporting image retrieval framework with rule base system" Knowledge-Based Systems (2010) .

[9] D. He and D. Wu "Enhancing query translation with relevance feedback in translingual information retrieval" Information Processing and Management 47, pp. 1–17, 2010

[10] H. Zhao and W. Jia, "An Adaptive Fuzzy Clustering Method for Query-by-Multiple-Example Image Retrieval" Third

International IEEE Conference on Signal-Image Technologies and Internet-Based System, pp. 997-1004, 2007

[11] H. Imran and A. Sharan "THESAURUS AND QUERY EXPANSION" International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, pp. 89-97, 2009

[12] Z. Zhou, Y. Tian, Y. Li, T. Huang and Wen Gao "Large-Scale Cross-Media Retrieval of WikipediaMM Images with Textual and Visual Query Expansion" Springer-Verlag Berlin Heidelberg, pp. 763-770, 2009

[13] D. Kim and S. Yu "A new region filtering and region weighting approach to relevance feedback in content-based image retrieval" The Journal of Systems and Software 81, pp. 1525–1538, 2008.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa "An evaluation of descriptors for large-scale image retrieval from sketched feature lines" Computers & Graphics 34, pp. 482–498, 2010

[15] L. Cinque, G. Ciocca, S. Levialdi, A. Pelocano and R. Schettini "Color-based image retrieval using spatial-chromatic histograms" Image and computing 19, pp. 979-986, 2001.

[16] Zhiguo Gong, MaybinMuyeba, JingzhiGuo, "Business information query expansion through semantic network", pp. 1-22, Enterprise Information Systems Volume 4 Issue 1, February 2010

[17] Yokoyama, A. Klyuev, V., "Search Engine Query Expansion Using Japanese WordNet", Human-Centric Computing (HumanCom), pp. 1-5, 3rd IEEE international conference, August, 2010

[18] Ming-Hung Hsu, Ming-Feng Tsai, Hsin-Hsi Chen, "Combining WordNet and ConceptNet for automatic query expansion: a learning approach", pp. 213-224, 4th Asia information retrieval conference on Information retrieval technology, AIRS'08, 2008

[19] C. Fellbaum, WordNet: an electronic lexical database. Cambridge, Mass: MIT Press, 1998.

[20] H. Liu and P. Singh, "ConceptNet a practical commonsense reasoning tool-kit," BT Technology Journal, vol. 22, no. 4, pp. 211–226, 2004.

[21] Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison by Ming-Hung Hsu, Tsai and Hsin-His Chen Springer 2006.

[22] B. Russell, A. Torralba and W. Freeman The open annotation tool http://labelme.csail.mit.edu/ Computer Science and Artificial Intelligence Laboratory, University MIT, 2005.

[23] Silva, J. Souza and K. Santos "Dependence Among Terms in Vector Space Model" Proceedings of the International Database Engineering and Applications Symposium, IEEE, 2004

[24] M. Flicker, H. Sawhney, W. Niblack and J. Ashley et al "Query By Image and Video Content: The QBIC System" 11th Annual Computer Security Application Conference, IEEE, pp. 23-32, 1995.

[25] W. Ma and B. Manjunath "NeTra: A toolbox for navigating large image databases" Multimedia Systems 7, pp. 184–198 Springer-Verlag 1999.

[26] J. Smith and S. Chang "VisualSEEk: a fully automated content-based image query system" ACM Multimedia 1996.

[27] J. Smith and S Chang "VisuallySearchingtheWeb for Content"pp. 12-20, IEEE Multimedia, 1997.

[28] Gupta "Visual Information Retrieval TechnologyAVirage Perspective" Virage Image Engine API Specification 1997.

[29] S. Chang, W. Chen, H. Meng, H. Sundaram and D. Zhong "VideoQ: An Automated Content Based Video Search System Using Visual Cues" ACM Multimedia, pp. 313-324, 1997

[30] M. Binderberger, S. Mehrotra, K. Chakrabarti and K. Porkaew " WebMARS: A Multimedia Search Engine" IEEE, 1999

[31] B. Luo, X. Wang and X. Tang "A World Wide Web Based Image Search Engine Using Text andImage Content Features" Proceedings of SPIE-IS&TElectronic Imaging, SPIE Vol. 5018, pp. 123-130, 2003

[32] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22, pp.1349–1380, 2000

[33] I. Silva, J. Souza and K. Santos "Dependence Among Terms in Vector Space Model" Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04), IEEE, 2004.

[34] X. Tai, F. Ren and K. Kita" An information retrieval model based on vector space method by supervised learning" Information Processing and Management 38, Elsevier Science, pp. 749–764, 2002.

Two-Layered Fuzzy Logic Based Frequency Modulation Controller for a Two-Area Thermal Reheat Power System Interconnected with an AC/DC Hybrid Transmission System

V.Adhimoorthy[1] and Dr.I.A.Chidambaram[2]

[1]Annamalai University, Department of Electrical Engineering
[2]Department of Electrical Engineering, Annamalai University,
[1]Assistant Professor, [2]Professor
adhisuganthi@gmail.com, driacdm@yahoo.com

Abstract: This paper presents new applications of logic technique for designing decentralized controllers for two-area interconnected thermal reheat power systems without and with high voltage direct current [HVDC] parallel link in a AC tie-line. The proposed two layered fuzzy controller, with the updated reference value of area control error [ACE] using pre-compensator ensures the ACE to zero with the inclusion of proportional plus Integral (PI) controllers. The Integral square error technique is adopted in optimizing the PI controller gains. When an AC power system is subjected to load disturbances, considerable frequency oscillations may result to system instability. So as to ensure the system stability, the power modulation control offered by HVDC link is enhanced to suppress the peak value of the transient frequency deviation. Simulation results show that the proposed two layered fuzzy logic controller is not only effective in damping out frequency oscillations, but also capable of alleviating the transient frequency swing caused by large load disturbance. Moreover, the output results prove that the present two layered fuzzy load frequency controller provides very good transient and steady state response compared to the fuzzy controller and conventional PI controller.

Keywords: Load-Frequency Control, Area control error, Integral squared error criterion, Fuzzy logic controller, Flexible AC Transmission system

1. Introduction

Load frequency control (LFC) is a very important issue in power system with an increasing demand for electric power and more complicated. Therefore the objective of LFC of a power system is to maintain the frequency of each area and tie-line power flow (in interconnected system) within specified tolerance by adjusting the new outputs of LFC generators so as to accommodate fluctuating load demand. A number of control schemes have been employed in the design of load frequency controllers [1] in order to achieve better dynamic performance. Among the various types of load frequency controllers the most widely conventional types used are the tie-line bias control and flat frequency control to achieve the above goals of LFC, both schemes are based on the classic controls which work on same function made up of the frequency and tie-line power deviations. Nevertheless these conventional control systems have been successful to some extent only [2]. This suggests the necessity of more advanced control strategies to be incorporated for better control. In this aspect if ensuring a better power quality intelligent controllers [2-8] have been replacing conventional controllers because of their fast and good dynamic response for load frequency control problems.

As the load demand increases tremendously, the power transmission over large distances to the remotely located load centres are forces to emerge into new plant for more and more effective and efficient control schemes for a better secured, reliable and stable system operation. This can be achieved by properly designed load-frequency control schemes i.e either by the proper selection of the controller or by incorporating efficient FACTS devices. [9-12]

In this paper the dynamic performance of two-area thermal reheat power system interconnected with AC tie-line and AC/DC hybrid tie-lines are considered and the PI controllers with various control schemes are designed and verified. In the AC/DC hybrid transmission system the HVDC link quickly starts the control system to suppress the peak value of transient frequency deviation hence a HVDC links is installed in parallel with an AC tie line in order to supply more to the area in need. In practical cases, the system parameters do not remain constant and continuously vary with changing operations.

Fuzzy logic controllers have received considerable interest in recent years. Fuzzy based methods are found to be very useful in the places where the solution to the mathematical formulations is complicated. Moreover, fuzzy logic controller often yields superior results to conventional control approaches [2-5]. The fuzzy logic based intelligent controllers are designed to facilitate the operation smooth and less oscillatory when system is subjected to load disturbances.

In this paper, the control scheme consists of two layers viz fuzzy pre-compensator and fuzzy PI controller. The purpose of the fuzzy pre-compensator is to modify the command signals to compensate for the overshoots and improve the steady state error. Fuzzy rules from the overall fuzzy rule vectors are used at the first layer, linear combination of independent fuzzy rules are used at the second layer. The two layer fuzzy system has less number of fuzzy rules as compared with the fuzzy logic system. The proposed two layered fuzzy logic controllers give better simulation results which is compared with the simulation results obtained using the fuzzy logic controllers and conventional controllers. Thus the two layered fuzzy PI controller enhances an efficient way of coping with imperfect information, offers flexibility in decision making processes

2. Application of AC/DC Hybrid Transmission System for the Proposed Work

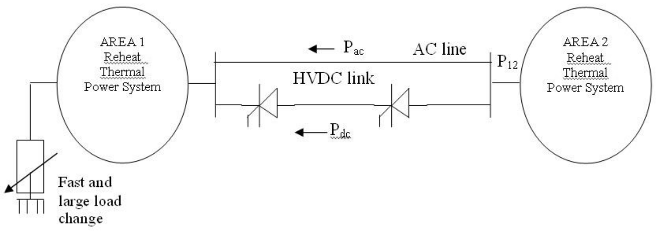


Figure 1. Two-area thermal reheat power system interconnected with an AC/DC hybrid transmission system

In the AC/DC hybrid transmission system the HVDC link consists mainly of a rectifier at the area 2 side, an inverter at the area 1 side and a DC transmission line apart from AC transmission line. In this system, It is assumed that, area 2 has supplied power P_{Ac} via only AC line to area 1. Next, there are installations of large loads with sudden charge in area 1. Therefore, the demand of electric power in area 1 increases further more and these large load change causes a serious problem of frequency oscillations in area 1. This implies that the capabilities of frequency control of governors in area 1 are not capable of stabilizing the frequency oscillations. On the other hand area 2 has enough frequency control capability to compensate for area 1. Therefore area 2 has an HVDC link installed in parallel with an AC tie-line in order to supply more power to area 1[13, 14].

In addition, area 2 offers stabilization of frequency oscillations to area 1 via HVDC link. The DC tie-line power modulation is capable of stabilizing frequency oscillations of area1 by complimentarily utilizing the control capability of area 2. According to the proposed control, the power system that has large capability of frequency control is able to offer service of frequency stabilization for other interconnected areas so that they do not have insufficient capabilities. The proposed control strategy can also be expected as a new ancillary service for stabilizing frequency oscillations.

The frequency modulation controller is modeled as a proportional controller of active power. It should be noted that the power modulation output of HVDC link [ΔP_{dc}], acting positively on area, reacts negatively an another area in an interconnected system ΔP_{dc},, therefore flow into both areas with different sign[+, -] simultaneously the time constant T_{dc} of proportional controller is set appropriately at 0.5[sec] in the simulation study.

3. Problem formulation

3.1 Model 1:
The linearized mathematical model of two – area thermal reheat power system is shown in figure 2 is represented by state variable equation as follows
The state space equations are

$$\dot{\mathrm{x}} = \mathrm{Ax} + \mathrm{B}u + \Gamma d \qquad (1)$$

$$y = \mathrm{Cx} \qquad (2)$$

where $\mathrm{x} = [\mathrm{x}_1^{\mathrm{T}}, \Delta \mathrm{p}_{\mathrm{ei}} \ldots \mathrm{x}_{(\mathrm{N-1})}^{\mathrm{T}}, \Delta \mathrm{p}_{\mathrm{e(N-1)}} \ldots \mathrm{x}_{\mathrm{N}}^{\mathrm{T}}]^{\mathrm{T}}$, n - state vector

$$\boldsymbol{n} = \sum_{i=1}^{N} n_i + (N-1)$$

$$u = [u_1, \ldots u_N]^T = [\Delta P_{C1} \ldots P_{CN}]^T, N \text{ -}$$

Control input vector

$$d = [d_1, \ldots d_N]^T = [\Delta P_{D1} \ldots P_{DN}]^T, N \text{ -}$$

Disturbance input vector

$$\mathrm{y} = [y_1 \ldots y_N]^T, \ \ 2N$$ - Measurable output vector

where A is system matrix, B is the input distribution matrix, Γ is the disturbance distribution matrix, C is the control output distribution matrix, x is the state vector, u is the control vector and d is the disturbance vector consisting of load changes.

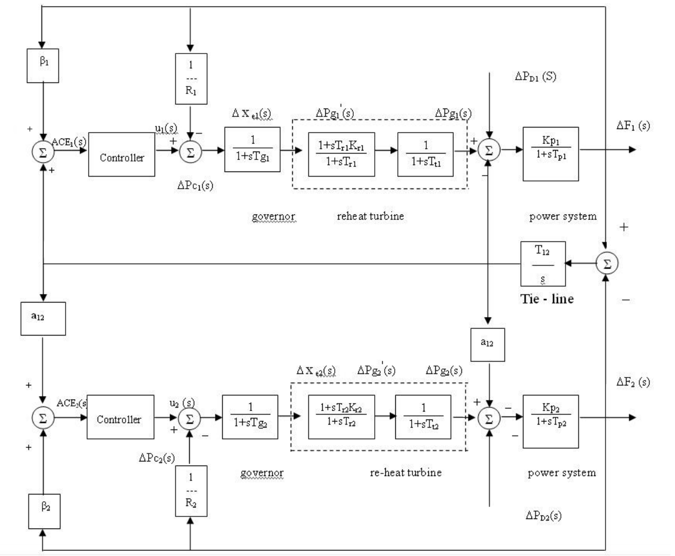


Figure 2. Block diagram of a two – area interconnected power system with reheat turbines

3.2 State Space Model

From the transfer function model shown in Fig.2 the following equation can be written by inspection.

$$\Delta \dot{F}_1 = \frac{k_{p1}}{T_{p1}} \left(\Delta P_{G1} - \Delta P_{D1} - \Delta P_{tie,1} \right) - \frac{\Delta F_1}{T_{p1}} \qquad (3)$$

$$\Delta \dot{P}_{G1} = -\frac{1}{T_{r1}}\Delta P_{G1} + \left[\frac{1}{T_{r1}} - \frac{k_{r1}}{T_{t1}}\right]\Delta P'_{G1} + \frac{k_{r1}}{T_{t1}}\Delta X_{E1} \quad (4)$$

$$\Delta \dot{P}'_{G1} = -\frac{1}{T_{t1}}\Delta X_{E1} - \frac{1}{T_{t1}}\Delta P'_{G1} \quad (5)$$

$$\Delta \dot{X}_{E1} = -\frac{1}{T_{g1}}\Delta X_{E1} + \frac{1}{T_{g1}}\Delta P_{c1} - \frac{1}{T_{g1} R_1}\Delta F_1 \quad (6)$$

$$\Delta \dot{P}_{tie,1} = T_{12}\left(\Delta F_1 - \Delta F_2\right) \quad (7)$$

$$\Delta \dot{F}_2 = \frac{k_{p2}}{T_{p2}}\left(\Delta P_{G2} - \Delta P_{D2} - a_{12}\Delta P_{tie,1}\right) - \frac{\Delta F_2}{T_{p2}} \quad (8)$$

$$\Delta \dot{P}_{G2} = -\frac{1}{T_{r2}}\Delta P_{G2} + \left[\frac{1}{T_{r2}} - \frac{k_{r2}}{T_{t2}}\right]\Delta P'_{G2} + \frac{k_{r2}}{T_{t2}}\Delta X_{E2} \quad (9)$$

$$\Delta \dot{P}'_{G2} = -\frac{1}{T_{t2}}\Delta X_{E2} - \frac{1}{T_{t2}}\Delta P'_{G2} \quad (10)$$

$$\Delta \dot{X}_{E2} = -\frac{1}{T_{g2}}\Delta X_{E2} + \frac{1}{T_{g2}}\Delta P_{c2} - \frac{1}{T_{g2} R_2}\Delta F_2 \quad (11)$$

3.3 Model 2:

The linearized mathematical model of two area thermal reheat interconnected system with HVDC links shown in Figure. 3 and state variable equations as follows

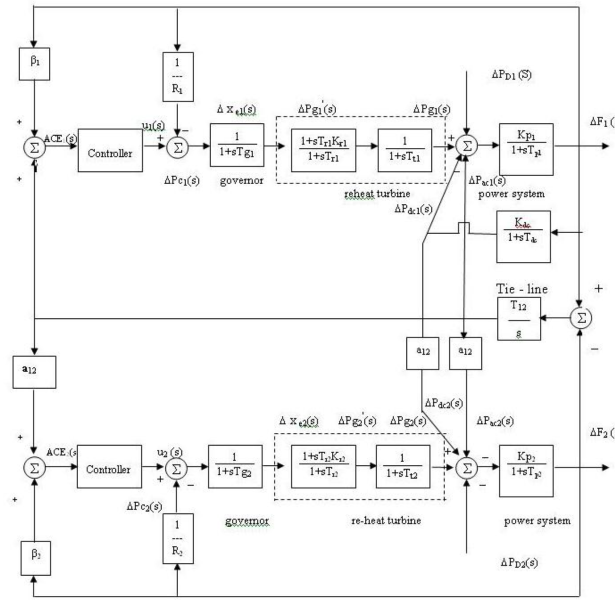


Figure 3. Block diagram of a two – area reheat thermal power system interconnected with AC/DC hybrid transmission system

$$\Delta \dot{F}_1 = \frac{K_{p1}}{T_{p1}}\left(\Delta P_{G1} - \Delta P_{D1} - \Delta p_{ac1} - \Delta p_{dc1}\right) - \frac{\Delta F_1}{T_{p1}} \quad (12)$$

$$\Delta \dot{P}_{G1} = -\frac{1}{T_{r1}}\Delta P_{G1} + \left[\frac{1}{T_{r1}} - \frac{k_{r1}}{T_{t1}}\right]\Delta P'_{G1} + \frac{K_{r1}}{T_{t1}}\Delta X_{E1} \quad (13)$$

$$\Delta \dot{P}'_{G1} = -\frac{1}{T_{t1}}\Delta X_{E1} - \frac{1}{T_{t1}}\Delta P'_{G1} \quad (14)$$

$$\Delta \dot{X}_{E1} = -\frac{1}{T_{g1}}\Delta X_{E1} + \frac{1}{T_{g1}}\Delta P_{c1} - \frac{1}{T_{g1} R_1}\Delta F_1 \quad (15)$$

$$ACE_1 = \beta_1\, \Delta F_1 + \Delta P_{ac1} + \Delta P_{dc1} \quad (16)$$

$$\Delta \dot{P}_{ac1} = 2\pi T_{12}\left(\Delta F_1 - \Delta F_2\right) \quad (17)$$

$$\Delta P_{dc1}(s) = \frac{K_{dc}}{T_{dc}}\Delta F_1 - \frac{1}{T_{dc}}\Delta_{pdc_1} \quad (18)$$

$$\Delta \dot{F}_2 = \frac{K_{p2}}{T_{p2}}\left(\Delta P_{G2} - \Delta P_{D2} - \Delta P_{ac2} - \Delta P_{dc2}\right) - \frac{\Delta F_2}{T_{p2}} \quad (19)$$

$$= \frac{k_{p2}}{T_{p2}}\left(\Delta P_{G2} - \Delta P_{D2} - a_{12}\Delta P_{ac1} - a_{12}\Delta P_{dc1}\right) - \frac{\Delta F_2}{T_{p2}} \quad (20)$$

$$\Delta \dot{P}_{G2} = -\frac{1}{T_{r2}}\Delta P_{G2} + \left[\frac{1}{T_{r2}} - \frac{k_{r2}}{T_{t2}}\right]\Delta P'_{G2} + \frac{K_{r2}}{T_{t2}}\Delta X_{E2} \quad (21)$$

$$\Delta \dot{P}'_{G2} = -\frac{1}{T_{t2}}\Delta X_{E2} - \frac{1}{T_{t2}}\Delta P'_{G2} \quad (22)$$

$$\Delta \dot{X}_{E2} = -\frac{1}{T_{g2}}\Delta X_{E2} + \frac{1}{T_{g2}}\Delta P_{c2} - \frac{1}{T_{g2} R_2}\Delta F_2 \quad (23)$$

3.4 Integral Squared Error Criterion

In order to ensure zero steady state error condition an integral controller may suitability designed for the augmented system. To incorporate the integral function in the controller, the system equations (1) and (2) are augmented with new state variables defined as the integral of $ACE_i \left(\int v_i dt\right), i = 1,2,...N$.

The augmented system of the order $(N+n)$ may be described as

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u + \bar{\Gamma}d \quad (24)$$

$$\bar{x} = \begin{bmatrix} \int vdt \\ x \end{bmatrix} \begin{matrix} \}N \\ \}n \end{matrix} \quad (25)$$

Where

$$\bar{A} = \begin{bmatrix} 0 & C \\ 0 & A \end{bmatrix} \bar{B} = \begin{bmatrix} 0 \\ B \end{bmatrix} \text{ and } \bar{\Gamma} = \begin{bmatrix} 0 \\ \Gamma \end{bmatrix}$$

The problem now is to design the decentralized feedback control law

$$u_i = -k_i^T \bar{y}_i \qquad i = 1,2.\dots.,N \quad (26)$$

The control law equation may be written in-terms of v_i as

$$u_i = -k_i \int v_i dt \qquad i = 1.2.\dots.,N \quad (27)$$

where k_i is the integral feedback gain vector.

Controllers designed on the basis of ISE criterion ensure reduction of rise time to limit the effect of large initial errors, reduction of peak overshoot and reduction of settling time to limit the effect of small errors lasting for a long time [20]. Further, this criterion is often of practical significance because of the minimization of control effort. The conventional decentralized optimum proportional plus integral controllers are designed using output feedback for the above mentioned case studies.

The following quadratic performance index is considered to obtain the optimum decentralized controller output feedback proportional plus integral gains for the interconnected two-area (identical areas) thermal reheat power system $\left(k_{1i} = \cdots = k_{2i}\right)$.

$$J_i = \int_0^t \left(\mathrm{x}_{ei}^T W_i \, \mathrm{x}_{ei}\right) dt \qquad i = 1,2 \tag{28}$$

Where $W_i = diag\{w_{i1,} w_{i2}\}$ and $\mathrm{x}_{ei}{}^T = [\Delta f_i, \Delta \dot{p}_{ei}]$

w_{i1} and w_{i2} are weighting factors for the frequency deviation and tie–line power deviation respectively of area i.

4. Design of Fuzzy Logic Systems

Fuzzy logic systems belong to the category of computational intelligence technique One advantage of the fuzzy logic over the other forms of knowledge-based controllers lies in the interpolative nature of the fuzzy control rules. The overlapping fuzzy antecedents to the control rules provide transitions between the control actions of different rules. Because of this interpolative quality, fuzzy controllers usually require far fewer rules than other knowledge-based controllers [7,8].

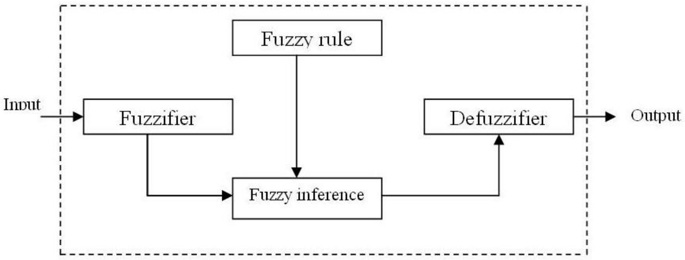


Figure 4. Block diagram of fuzzy logic controller

A fuzzy system knowledge base consists of a fuzzy if then rules and membership functions characterizing the fuzzy sets. The block diagram and architecture of fuzzy logic controller is shown in fig4. Membership function (MF) specifies the degree to which a given input belongs to a set. Here triangular membership function have been used to explore best dynamic responses namely negative big(NB), negative small(NS) , zero(ZE), positive small (PS), positive big(PB). Fuzzy rules are conditional statement that specifies the relationship among fuzzy variables. These rules help to describe the control action in quantitative terms and have been obtained by examining the output response to the corresponding inputs to the fuzzy controllers. Defuzzification, to obtain crisp value of FLC output is done by centre of area method. The fuzzy rules are designed as shown in table 1.

Table 1. Fuzzy Logic Rules For LFC

ACE $\dot{ACE}$	**NB**	**NS**	**Z**	**PS**	**PB**
NB	NB	NB	NS	NS	ZE
NS	NB	NB	NS	ZE	ZE
Z	NS	NS	ZE	PS	PS
PS	ZE	NS	PS	PS	PB
PB	ZE	ZE	PS	PB	PB

5. Two Layered Fuzzy Logic Controller

The aim of introducing two layered fuzzy logic controller [15] is to eliminate the steady state error and improve the performance of the output response of the system under study. The proposed control scheme is shown in Fig. 5. The controller consists of two "layers": a fuzzy pre-compensator and a usual fuzzy PI controller. The error e(k) and change of error Δe(k) are the inputs to the precompensator. The output of the pre-compensator is μ(k) The PI Controller is usually implemented as follows:

$$u(k) = k_p e(k) + TK_i \sum_{n=0}^{k} e(n) \tag{29}$$

Where $e(k) = y(k) - y_r(k)$ and Δ e(K) = e (k) - e(k-1)

The controller output, process output and the set point are denoted as u, y and y_r respectively. Experience-based tuning method - Ziegler-Nichols method which widely adopted [16] requires a close attention since the process has to be operated near instability to measure the ultimate gain and period. This tuning technique may fail to tune the process with relatively large dead time [16]. In order to improve the performance of PI tuning a number of attempts have been made which can be categorized into two groups: Set point modification and gain modification.
The set point modification introduces new error terms

$$e_p = y_r(k) F_p(e, \Delta e) - y(k) \tag{30}$$

$$e_i = y_r(k) F_i(e, \Delta e) - y(k) \tag{31}$$

The corresponding control law is given by, Where F_p , F_i

are non linear functions of e and Δe.

$$u(k) = k_p e_p(k) + TK_i \sum_{n=0}^{k} e_i(n) \tag{32}$$

As a special case, one would like to modify the set point only in proportional terms. This implies F_p =β ; F_i =1 set point weight [17]

$$\therefore \text{ U(k)= } K_p\{\beta y_r(k) - y(k)\} + TK_i \sum_{n=0}^{k} e(n) \tag{33}$$

or

$$U(k) = K_p e'(k) + TK_i \sum_{n=0}^{k} e'(n)$$

The pre-compensation scheme [17, 18] is easy to implement in practice, since the existing PI control can be used without modification in conjunction with the fuzzy pre-compensator as shown in Figure. 5(a).

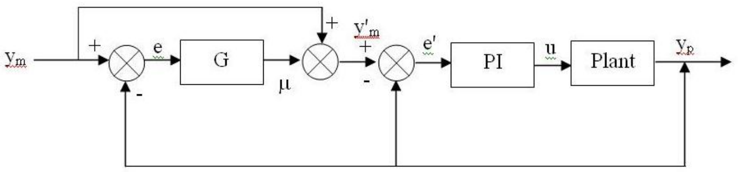


Figure 5. Basic structure of fuzzy pre-compensated PI controller

The procedure of rule generation consists of two parts (i) learning of initial rules which determines the linguistic values of the consequent variables. (ii) fine tuning adjusts the membership function of the rules obtained by the previous step. The structure of the pre-compensation rule is written as If e is L_e, and Δe is LΔe then C is L_c where L_e, ΔL_e and L_c are linguistic values of e, Δe, c respectively.

Each fuzzy variable is assumed to take 5 linguistic values Le, LΔe, or Lc = {NB, NS, ZE, PS, PB} this leads to fuzzy rules, if the rule base is complete.

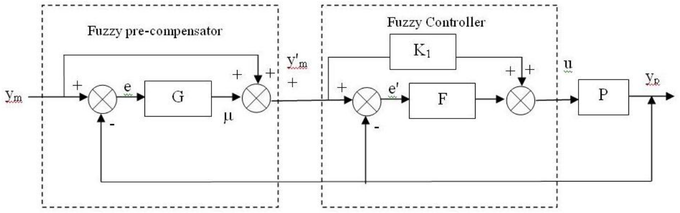


Figure 5 (a). Proposed two layered fuzzy logic controller

The dynamics of overall system is than described by following equations

$$e(k) = ym(k) - yp(k) \tag{34}$$

$$\Delta e(k) = e(k) - e(k-1) \tag{35}$$

$$\mu(k) = G[e(k), \Delta e(k)] \tag{36}$$

Where μ(k) is a compensating term which is generated using a fuzzy logic scheme

$$y'_m(k) = y_m(k) + \mu(k) \tag{37}$$

$$e'(k) = y'_m(k) + y_p(k) \tag{38}$$

$$\Delta e'(k) = e'(k) - e'(k-1) \tag{39}$$

The proposed two layered FLC compensate these defects and gives fast responses without large overshoot and/or undershoot. Moreover to steady state error reduces to zero. The first layer fuzzy pre compensator is used to update and modify the reference value of the output signals to damp out oscillations. The fuzzy states of the input and output all are chosen to be equal in number and use the same linguistic descriptors as N = Negative, Z = Zero, P = Positive to design the fuzzy rules. The fuzzy logic rules for precompensator are presented in Table-2.

Table 2. Fuzzy Logic Rules for Precompensator

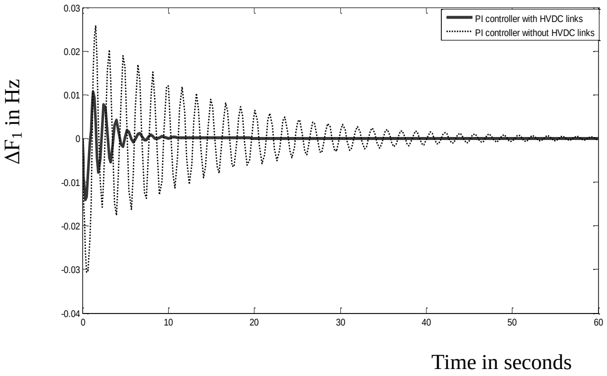
ACE $\dot{ACE}$	N	Z	P
N	N	N	N
Z	Z	Z	Z
P	Z	P	P

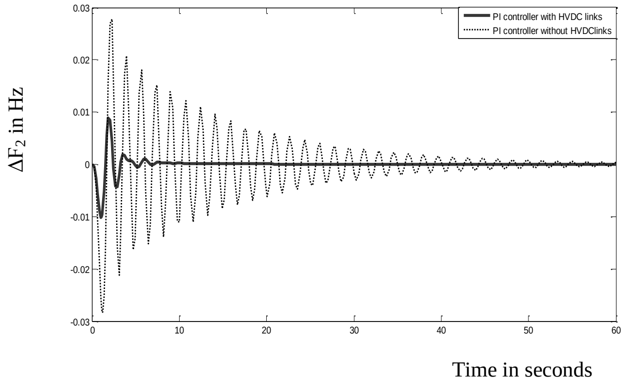
The second layer which is known as feedback fuzzy logic control reduces the steady state error to zero. The output of the FLC is given by

$$u(k) = K_1 y'_m(k) + F[e'(k), \Delta e'(k)] \tag{40}$$

6. Simulations Results and Observations

The optimal gains of the conventional PI controller are determined on the basis of Integral Squared Error (ISE) technique by minimizing the quadratic performance index. This controller is implemented in the interconnected two area power systems without and with HVDC link for 1% step load disturbance in area 1. The conventional optimum gain values are found to be K_{P1}=K_{P2}=1.3 and K_{I1}=K_{I2}=0.08 for the two area interconnected power system without HVDC link, K_{p1}=K_{p2}=2.1 and K_{i1}=K_{i2}=0.25 with HVDC link. Moreover the fuzzy logic and two layered fuzzy logic controller are designed and implemented in the interconnected two area power system without and with HVDC link for 1% step load disturbance in area 1. Simulations results are shown in fig 6(a) and 6(b). It is found that the proposed two layered fuzzy logic controller has less over/ undershoots and ensures faster settling time and improvement in stability as compared with fuzzy logic controller and conventional PI controller.





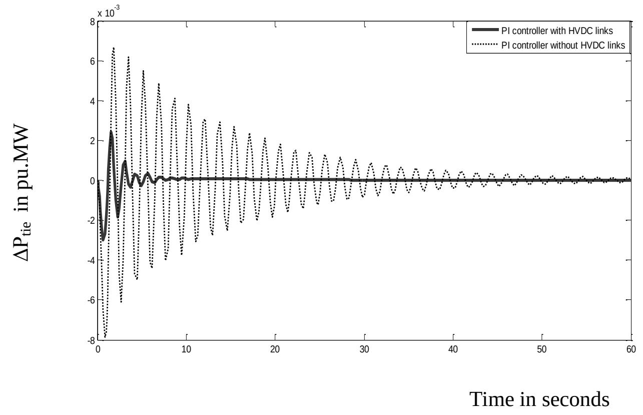


Figure 6(a). Dynamic responses of the frequency deviation and tie-line power deviations for two area thermal reheat power system with and without HVDC links considering 0.01 pu MW step load disturbance in area 1using PI controllers

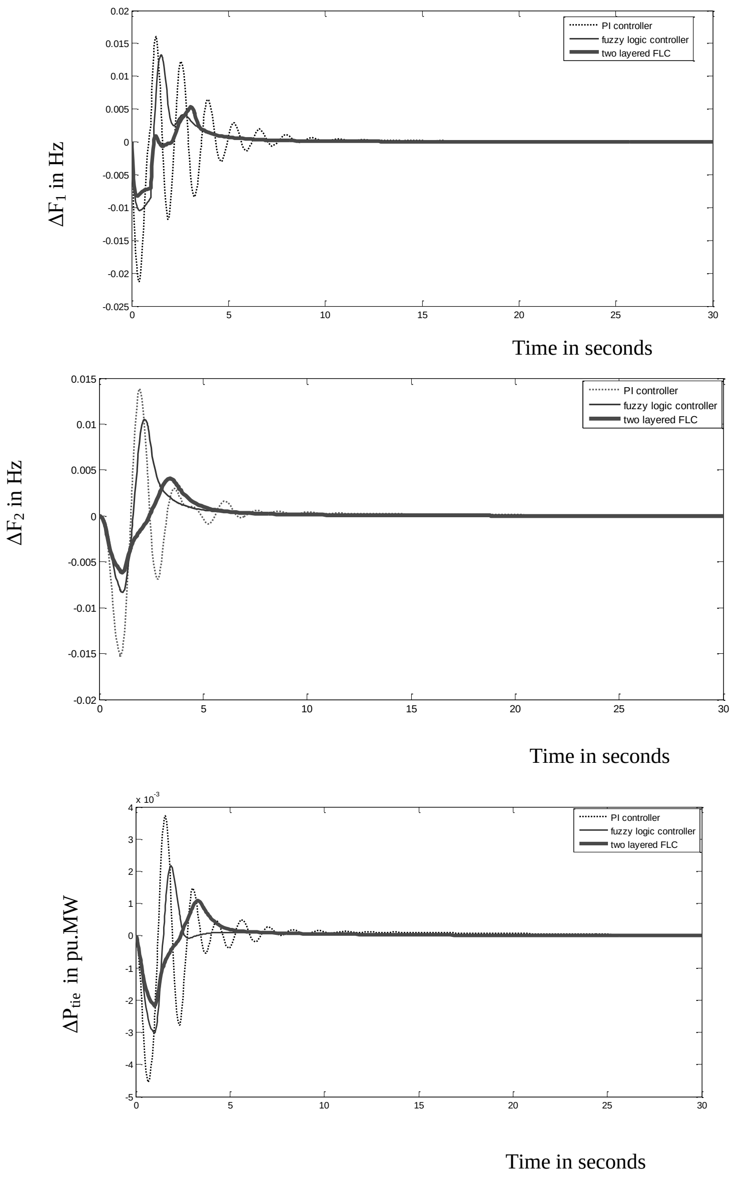


Figure 6(b). Comparison of the dynamic responses of the frequency deviations and tie line deviation for two area thermal reheat power system with HVDC links considering 0.01 pu.MW step load disturbance in area 1.

Conclusion:

In this paper, the conventional PI controllers are designed and implemented in a two-area thermal reheat power system interconnected with AC tie-line and with AC/DC hybrid tie-lines. From the dynamic response reveals that the two-area reheat power system when interconnected with AC/DC hybrid tie-line ensures far for better transient performance and faster settling time than that of the system with ac tie-lines. Moreover, the fuzzy PI controller and two layered fuzzy PI controllers were designed and implemented in the two area thermal reheat power system interconnected with ac/dc hybrid tie-line. The Dynamic responses of the system with these controllers are compared and it is found that the two layered fuzzy PI controllers are found to be the best among the three controllers as the two layered fuzzy PI controller exhibits very less frequency oscillations, very less tie-line power deviations and also very less control effort.

Acknowledgement:

The authors wish to thank the authorities of Annamalai University, Annamalainagar, Tamilnadu, India for the facilities provided to prepare this paper. The authors also wish to thank Dr. M.M.T.Ansari, Associate Professor, Department of Electrical Engg., Annamalai University, for the valuable suggestions in carrying out this paper

References:

[1]. Shayeghi.H, Shayanfar.H.A, Jalili.A, “Load-frequency control strategies: A state of the art survey for the researcher” Energy Conversion and Management, Vol. 50(2), pp.344-353, 2009.

[2]. C.C. Lee, “Fuzzy Logic in control systems Fuzzy logic controller, Part-1”, IEEE Transactions on System, Man and Cybernetics, Vol. 20(2), pp.419-435, 1990.

[3]. C.S. Indulkar and Baldevraj, “Application of fuzzy Controller to automatic generation Control”, Electric Machines and Power Systems, Vol. 23, pp.209-220, 1995.

[4]. Zadeh. L.A., “Fuzzy Algorithm–Information and Control”, Vol. 12, pp.94 – 102, 1969.

[5]. M.K.EL-Sherbiny, G.EL-Saady, Ali M. Yousef, “Efficient fuzzy logic load frequency controller”, Energy Conversion and Management Vol.43, pp.1853-1863, 2002.

[6]. H.D.Mathur, H.V.Manjunath, “Study of dynamic performance of thermal units with asynchronous tie-lines using fuzzy based controller”, Journal of Electrical Systems, Vol. 3, pp.124- 130, 2007.

[7]. J.K.Kim, J.H.Park, S.W.Lee, K.P.Chong, “A Two-Layered Fuzzy Logic Controller for Systems with Deadzones”. IEEE Transactions on Industrial Electronics, Vol-41, No. 2 , pp.155-166 , 1994.

[8]. Ahmed Rubaai, Abdul R.Ofoli, “Multilayer fuzzy controller for power networks”, IEEE Transaction on Industrial Applications, Vol. 40, No.6. pp. 1521-1528, 2004.

[9]. Issarachai Ngamroo, "A Stabilization of Frequency Oscillations in a Parallel AC-DC interconnected Power System via an HVDC Link", Science Asia, Vol. 28, pp.173-180, 2002

[10]. Bhamidipati.S, Kumar.A., 'Load frequency control of an inter-connected system with DC tie-lines and AC-DC parallel tie-lines', Proceedings of the Twenty-Second Annual North Digital Object Identifier 10.1109/NAPS.1990.151393, pp.390-395, 1990

[11]. I.A. Chidambaram, S.Selvakumaran and S.Velusami, "Dual Mode Controller for Load-Frequency Control of an interconnected Power System with AC / DC Tie-lines", National Conference on Recent trends in Industrial Electronics, Drives and Controls, Annamalai University, Annamalainagar, pp.168-173, 2003

[12]. Ibrahim, P.Kumar, S.Khatoon, "Effect of parameter uncertainties on dynamic performance of an interconnected power system with AC/DC links", International Journal of Power and Energy Systems, pp.3203-3259, 2005.

[13]. Kolonskill, T.V, "Emerging control of synchronous and non synchronous tie lines", Power Eng (N.Y English Translation), Vol / issue: 18:6, pp.141-143, 1980

[14]. Prabhatkumar, Ibrahim, "Dynamic Performance Evaluation of two-area interconnected power systems: a comparative study", IE(I) Journal-EL, pp.199-209,1998

[15]. Moon.G.Joo, "A method of converting conventional fuzzy logic system to two layered hierarchical fuzzy system" IEEE International Conference on Fuzzy Systems,pp.1357-1362, 2003

[16]. CC. Hang, K.J. Astrom and WK. Ho, "Refinements of the Ziegler – Nichols tuning formula", IEEE Proceeding, Part D, Vol.138 (2), pp.111- 118, 1991

[17]. KH. Kim, K.C. Kim, E.K.P. Chong, "Fuzzy pre-compensation of PID Controller", IEEE Transactions on Control Systems Technology, Vol. 2, pp.406-411,1994

[18]. J.K. Kim, J.H. Park, SW Lee and KEP Chong, "Fuzzy pre-compensation of PD controller for system with dead zones", Journal on Intelligent fuzzy system, Vol. 1, pp.125-133,1993

[19]. Hyun-Joon Cho, K Wang-Bo Cho, Bo-Hyeun Wang, "Fuzzy PID hybrid control: Automatic rule generation using genetic algorithms", Fuzzy Sets and Systems, Vol. 92, pp.305- 316,1997

[20]. Ogata Katsuhiko, Modern Control Engineering, New Delhi, India, Prentice Hall of India private Limited, 1986.

Appendix:

(i) Data for Thermal Power System with Reheat Turbines [11].

f^0 = 60 Hz, PR_1 = PR_2 = 2000 MW, $K_{p1}=K_{p2}$= 120 Hz / pu.MW, T_{pS1} = T_{pS2} = 20 sec, T_{t1} = T_{t2} = 0.3 sec, $T_{g1}=T_{g2}$= 0.08 sec, $K_{r1}=K_{r2}$= 0.5, T_{r1} = T_{r2} = 10 sec, R1 = R2 = 2.4 Hz/p.u MW, $\beta_1=\beta_2$ = 0.425 pu.MW/Hz, ΔP_{D1} = 0.01 p.u MW, T = 2 sec (Normal sampling rate),

(ii) Data for AC link [14]

P_{tmax} 200 MW; T_{12}=0.545 pu.MW/Hz

(iii) Data for DC link [14]

K_{dc}=1.0; T_{dc}= 0.5 sec

Author Biographies:

V.Adhimoorthy (1974) received Bachelor of Engineering in Electrical and Electronics Engineering (2002), Master of Engineering in Power System Engineering (2008) and he is working as Assistant Professor in the Department of Electrical Engineering, Annamalai University He is currently pursuing Ph.D degree in Electrical Engineering at Annamalai University, Annamalainagar. His research interests are in Power Systems, Control Systems, Electrical Measurements. (Electrical Measurements Laboratory, Department of Electrical Engineering, Annamalai University, Annamalainagar-608002, Tamilnadu, India, adhisuganthi@gmail.com

I.A.Chidambaram (1966) received Bachelor of Engineering in Electrical and Electronics Engineering (1987), Master of Engineering in Power System Engineering (1992) and Ph.D in Electrical Engineering (2007) from Annamalai University, Annamalainagar. During 1988 - 1993 he was working as Lecturer in the Department of Electrical Engineering, Annamalai University and from 2007 he is working as Professor in the Department of Electrical Engineering, Annamalai University, Annamalainagar. He is a member of ISTE and ISCA. His research interests are in Power Systems, Electrical Measurements and Controls. (Electrical Measurements Laboratory, Department of Electrical Engineering, Annamalai University, Annamalainagar – 608002, Tamilnadu, India, Tel: - 91-04144-238501, Fax: -91-04144-238275) **driacdm@yahoo.com**,

Effect of Radio Platform Noise on Existing MAC Protocols in use in Ad Hoc Wireless Networks and its Solution

Prabir Banerjee[*], D.K.Basu[+], M.Nasipuri[+]
[*]ECE Department,Heritage Institute of Technology,Kolkata
[+]Computer Science & Engineering Department, Jadavpur University
prabir.banerjee@gmail.com

Abstract: *Ad Hoc wireless networks represent a communication paradigm that is relatively new. It has the potential to provide ubiquitous communication and this is the reason why it is generating such keen interest in the world of communication. One of the key factors for the success of Ad Hoc wireless networks is smooth interaction between the medium access control (MAC) layer and the physical layer. In this paper, we have modified the correlation between ALOHA, an established MAC protocol and the noise model of the physical layer for a circuit switched network. This modified model will help to estimate the maximum effective transport capacity more correctly in-between the nodes of Ad Hoc wireless networks. This model will give more precise result since it has also taken into account the noise generated by the node-radio platform.*

Key-words: *ALOHA, Inter-node interference (INI), Medium access control (MAC), Signal-to-noise ratio (SNR), Transport capacity*

1. Introduction

Ad Hoc wireless networks promise newer and ubiquitous communication anywhere since they are independent of any infrastructure and so-called base-stations. At the same time, such networks face quite a few challenges with respect to the wireless networks with defined bases and infrastructures.

One of the biggest challenges is to operate with low capacity battery and omni-directional antenna. As a result, the signal –to-noise ratio (SNR) at any receiver node cannot be increased by classical means like increasing the transmitter power or making the signal directive. Therefore, the noise element at the receiver assumes far greater significance.

We know that the noise element at any wireless receiver is taken to be the addition of two factors- one is the link noise and the other is the thermal noise both of which are functions of the bandwidth of the receiver. If we consider the overall construction of any node-radio, it consists of three modules namely, (i) receiver and transmitter block; (ii) power supply block and (iii) the control function block.

The control block for any present day radio plays the most important role- it provides functions like channel selection, phase locked loop(PLL) control, switching of active mode and sleep mode of the microcontroller etc. and ,therefore, this block has become the integral part of even radios used in the nodes of the Ad Hoc wireless networks.

Hence, it is but logical that this block has to be considered for any possible additional noise factor while the SNR of the received signal is considered.

Now, let us examine the data transfer mechanism in an Ad Hoc wireless network so that the possible degradation due to fundamental performance limits can be explained.

The concept of transport capacity has been introduced to quantify the achievable transmission of information in the Ad Hoc wireless networks and it involves two parameters- the rate of data transfer and the distance between the source node and the destination node.

In [1], the authors have computed the transport capacity of stationary wireless networks. It also assumes that inter-node interference (INI) will be nil for the given Ad Hoc wireless network. This is possible only when the SNR for both the source and destination nodes is above the threshold value for the network. Here, SNR considers both interference and thermal noise.

This approach [1], does not give a clear picture about the influence of physical layer characteristics and of the medium access control (MAC) protocol on the achievable performance.

The paper [2] shows how physical layer and MAC layer are interrelated. We have applied the results to establish that our suggested modification will make the predicted transport capacity more correct and realistic in Ad Hoc wireless networks.

The remainder of this paper is structured as follows. In section 2, a model for circuit switched Ad Hoc wireless networks is described. Section 3 details the background logic and modification of the model. In section 4, ALOHA MAC protocol and its operation with respect to new model is explained. Section 5 draws the conclusion.

2. Ad Hoc wireless networks and circuit switching

If we consider any Ad Hoc wireless network with reasonably high node spatial density, it is but natural that inter-node interference will occur most of the time.

Now, concept of effective transport capacity [1] is a very useful concept to compare different mechanisms in an Ad Hoc wireless network and it represents the rate-distance product actually carried by the network. The rate-distance is basically the product of the maximum data rate and maximum distance so that the product can be used as an indicator for reliable data transfer in the wireless network.

If the physical model is one of non-interference, then it is assumed that the SNR at the receiving node is above the threshold value for the Ad Hoc wireless network in spite of interference from other active nodes of the same network or other adjoining networks. In [1], it is also assumed that the network is a stationary one. Therefore, we have also assumed these conditions for our paper.

Different schemes like spread-spectrum technique can be employed to combat problem of interference but in this paper this angle has not been pursued and the attention is on the random access MAC scheme namely, ALOHA. Techniques like automatic repeat request(ARQ) have not been considered since such techniques are based on the concept of re-transmission of the data packets and for Ad Hoc wireless networks, energy conservation at the nodes is one of critical importance. Rather, we have considered a bit-level interference analysis using ALOHA MAC protocol.

The objective of any Ad Hoc wireless network is to reach data to the desired destination node from the originating source node – the route usually consists of a number of nodes giving rise to multi-hop communication. For a created route, the nodes involved can be thought of as forming a communication tube.

These tubes can bend if the nodes become mobile and the configuration of the tubes changes when the existing route does not work any longer necessitating drop or insert of nodes. The creation of a private path between the source and destination nodes resembles circuit switching [3].

In each of such tubes, there are gaps between consecutive packets to ensure that there is no INI. It is assumed in the rest of this paper that the packet transmission is Poisson distributed with parameter λ. It implies that the average inter-arrival rate between two consecutive packets is 1/ λ.

If R_b is the channel data-rate of any node in the Ad Hoc wireless network and L is the number of bits in a packet, the packet duration is given by (L/ $R_{b)}$. If (L/ $R_{b)}$ is sufficiently smaller than 1/ λ, the inter-arrival time, the packets transmitted in the two tubes may not overlap reducing significantly the inter-route interference. This idea of non interference applies to the ALOHA scheme.

3. The mathematical model

On the basis of the communication theoretic framework developed in [2], we have considered a node distribution pattern in which the total of N nodes of the Ad Hoc wireless network are placed at the vertices of a square grid inside a circular area A.

The node spatial density, ρs, defined as the number of nodes per unit area is then (N/A). If r_L is the minimum inter-node distance, then:

$$r_L \approx 1/(\rho s)^{½} \quad \text{------(1)}$$

It is assumed subsequently that a multi-hop route is formed by a sequence of minimum length hops.

This is a very powerful strategy to minimize the end-to-end bit error rate (BER) keeping the transmission power level to a minimum [2] [4]. If BER at the end of a single hop is denoted by p_L and it is assumed that (i) each node of the network is a regenerative one and that (ii) the uncorrected errors accumulate, it is possible to show [2], that the BER at the end of the nth. link of a multi-hop route can be expressed as:

$$P_b^{(n)} \approx 1-(1-p_L)^n \quad \text{-------(2)}$$

The average BER can be calculated by evaluating (1) for an average number of hops. If it is assumed that the number of hops is uniformly distributed between 1 and $n_h^{max} = 2(N/\pi)^{½}$, the average number of hops becomes

$$n_h^{av} = [\,((N/\pi)^{½}\,].$$

Hence, from (1) and (2):

$$P_b^{av} = P_b^{(n_h^{av})} \approx 1-(1-p_L)^{[((N/\pi)^{½\,]}} \quad \text{---(3)}$$

Where, P_b^{av} is the average value of BER.

The link BER, p_L is directly related to the SNR at the destination node of the link and let it be denoted by $SNR_{L.}$ We are further assuming that that the transmitted signal is affected only by the free-space loss. Therefore, the received signal power at the end of a minimum length hop will be governed by Friis equation.

This received power $P_r^{(r_L)}$ can be expressed as:

$P_r^{(r_L)} = (\alpha .\ P_t\ /\ r_L^2) \approx \alpha .\ \rho s\ .\ P_t$ [using (1)]

$$= \frac{G_t \,.\, G_r \,.\, \lambda_c^2}{(4\pi)^2\ .\ L} \,.\, \rho s \,.\, P_t \text{ ---(4)}$$

where, P_t is the power transmitted from each node of the link in the Ad Hoc wireless network, λ_c is the wavelength of the carrier frequency, G_t and G_r are the gains of the transmitter and receiver antenna respectively and L is the loss factor of the medium.

Generally, two distinct cases are distinguished based on the presence or absence of INI – one is the ideal case in which the interference is totally absent and second one in which the INI is considered to be present. We have analysed the realistic or the second case in more detail and added an additional element in the model.

i) Ideal case – in this case, interference from other nodes is assumed to be absent so that the noise at the receiving node consists only of the thermal noise generated at the receiver. The link SNR can be expressed as

$$SNR_L^{noINI} = \frac{P_r^{(rL)}}{P_{thermal}} \text{ -----(5)}$$

where $P_{thermal}$ is the thermal noise power at the receiver.

ii) Realistic case(INI present) – in such a scenario, the interfering signals from other nodes of the Ad Hoc wireless network may be treated as additive white noise independent from the thermal noise of the receiving node. The SNR at the end of a minimum link length can be expressed as

$$SNR_L^{INI} = \frac{P_r^{(rL)}}{P_{thermal} + P_{INT}} \text{ ------(6)}$$

where P_{INT} is the interfering signal power.

3A. The modification to the SNR

For any new generation radio, the operation of the radio is controlled mostly by intelligent programs which run on fast micro-controllers. This control platform introduces significant noise in the radio particularly the receiver mixer section[..]. It may cause desensitization of radios at particular frequency thereby reducing the range or increasing the noise in the receiver. The resultant SNR should ,therefore, be expressed as:

$$SNR_L^{net} = \frac{P_r^{(rL)}}{P_{thermal} + P_{INT} + Ppf} \text{ ------(7)}$$

where, P_{pf} represents the platform noise power at the receiving node of the Ad Hoc wireless network.

We refer to full connectivity when at the end of the desired multi-hop communication route, the BER is lower than a maximum tolerable value. Since the link BER is a decreasing function of the SNR obtained in the link, it is but logical that SNR_L has to exceed a certain minimum value of SNR_L^{min} . It again depends on P_b^{max} and N.

Since, the platform noise power may change the SNR value of the link drastically, it is very important that this factor is removed through proper hardware design or taken care of in calculating the range.

4. ALOHA MAC Protocol

The basic principle of Aloha protocol is the provision that each node, without sensing the channel occupancy, transmits whenever it has information to transmit. An Ad Hoc wireless network with multi-hops can use this protocol easily.

It is seen that in the case of ALOHA protocol, the link SNR is a monotonically increasing function of the transmitted power and node spatial density and is lower than a maximum value SNR_L for ALOHA. If there is strong internal interference as already discussed in section 3, the SNR of the link degrades and increasing the transmitter power does not solve the problem of poor SNR for the link.

Moreover, since the restriction on transmitter power level for the nodes of Ad Hoc wireless networks is severe because of importance of energy conservation, it is not a solution at all.

In [2],[5] , a new method for bit-level analysis is proposed and the BER performance with ALOHA protocol is analyzed. We have taken part of the results to evaluate the effective transport capacity and to show how it is affected by the presence of strong interference from the controller and digital hardware of the radio control circuit.

It is possible to show that the interference power appearing in the SNR expression (6) can be expressed as :

$$P_{INT}^{aloha} \approx \alpha.\ \rho s\ .\ P_t\ (1 - e^{-\lambda D}{}_{p}\ )\ .\ K \text{ -----(8)}$$

Where D_p is the packet duration, given by L / R_b and K is a factor dependent on the maximum tier number in a square grid network. The $SNR_L^{aloha,max}$ can be written as :

$$SNR_L^{aloha,max} = \lim_{P_t,\ \rho s \to \infty} SNR_L^{aloha} \text{ --(9)}$$

For strong interfering power represented by P_{pf} in equation (7), this SNR equation for Ad Hoc wireless networks using ALOHA mac protocol will have a serious implication.

The $SNR_L^{aloha,max}$ will decrease and in some cases, it may become lower than the value SNR_L^{min} required for full connectivity of the Ad Hoc wireless network.

It will fail irrespective of the transmitted power level and node spatial density.

5. Solution

Such unpredictability of SNR value can be a serious problem for successful operation of Ad Hoc wireless networks. To prevent such situation, hardware solutions were suggested after exhaustive experiments in our paper [8].

The paper describes two possible solutions to eliminate the problem of radio frequency interference(RFI) from the radio platform.

The second of the solution is a better one and briefly, it has a built-in intelligent circuit which checks continuously the SNR of the radio of the receivers with the help of the radio signal strength indicator(RSSI) output from the intermediate frequency(IF) processor integrated circuit(IC). The micro-controller oscillator frequency is changed under program control till a satisfactory RSSI is achieved or in other words, a good and reliable SNR value is attained in the receiver.

When this is achieved, we can say that the interference factor as shown in equation (7) can be neglected for all practical purposes.

6. Conclusion

We have shown that the noise factors present in an wireless link are contributed by more than one element and one of those is the RFI from the platform of the radios of the Ad Hoc wireless network nodes. This can be the most damaging one since the desensitization of the receiver can be carrier frequency dependent and again ,in turn, the spurious frequency component could be any sub-harmonic of the controller clock frequency.

Hence, unless the root cause of this possible interference is eliminated in the hardware design stage, the performance of the Ad Hoc wireless networks can continue to be unreliable connectivity-wise and BER-wise.

Acknowledgement

The authors are indebted to CMATER and SRUVM projects of the Computer Science and Engineering(CSE) department, Jadavpur University for allowing them to use the laboratory and other facilities.

Prabir Banerjee is grateful to the authorities of Heritage Institute of Technology, Kolkata for allowing him to work on the research paper, and Dr.D.K.Basu to AICTE,New Delhi for providing him with an Emeritus fellowship.

References

[1] P.Gupta and P.R.Kumar, The capacity of wireless networks, IEEE Trans. Inform. Theory, vol.46,pp 388-404, March 2000.

[2] O.K.Tonguz and G.Ferrari, A Communication- Theoretic Framework for Ad Hoc wireless Networks, Carnegie Mellon University, ECE Dept., Tech. Rep., TR-043-2003, February 2003.

[3] D.Bertsekas and R.Gallager, Data Networks, Upper Saddle River, NJ, U.S.A., Prentice Hall, 1992, second edition.

[4] G.Ferrari and O.K.Tonguz, Performance of

circuit-switched ad hoc wireless networks with Aloha and PR-CSMA protocols, in IEEE Global Telecommun. Conf., San Francisco, USA, December 2003.

[5] G.Ferrari and O.K.Tonguz, A communication-theoretic framework for ad hoc wireless networks-Part II : MAC protocols and inter-node interference, 2003,submitted to IEEE Trans. On Commun..

[6] T.S.Rappaport, Wireless Communications, Principles and Practice, Prentice Hall,2002, second edition.

[7] J.G.Proakis, Digital Communication, 4th. ed.,Mcgraw Hill, 2001.

[8] Prabir Banerjee, D.K.Basu, M.Nasipuri, Mitigation of RFI in Ad Hoc Wireless Receiver Nodes, IJRTE Journal, vol.2, no.6,pp 115-120, November 2009

Bio-data of the authors

Prabir Banerjee obtained his Bachelor’s degree in B.E.Tel.E from the ETC department of Jadavpur University in 1980 and his M.E.Tel.E. degree in Computer Engineering from Jadavpur University in 1982.

He was with the research and development departments of Philips Telecommunication and then Simoco International till 2003 before joining the academia.

In the course of his career, he was involved in the development of synthesized PMR and repeater. He also developed Sel-call system for mobile networks in India and ATE for radios. He was involved in technology transfers of POCSAG pagers and state-of-the-art VHF/UHF radios. Prof. Banerjee also developed the label printing software for pagers.

Subsequently, he actively participated in designing HF networks in difficult terrains for police and para-military forces in India. He also redesigned part of HF sets to improve power and noise performance.

He has to his credit 12 published papers till date on ad-hoc wireless networks and is currently working on techniques to enhance the spectrum utilization efficiency.

Mita Nasipuri received her B.E.Tel.E., M.E.Tel.E., and Ph.D. (Engg.) degrees from Jadavpur University, in 1979, 1981 and 1990, respectively. Prof. Nasipuri has been a faculty member of J.U since 1987. Her current research interest includes image processing, pattern recognition, and multimedia systems. She is a senior member of the IEEE, U.S.A., Fellow of I.E (India) and W.B.A.S.T, Kolkata, India.

Dipak Kumar Basu received his B.E.Tel.E., M.E.Tel., and Ph.D. (Engg.) degrees from Jadavpur University, in 1964, 1966 and 1969 respectively. Prof. Basu has been a faculty member of J.U from 1968 to January 2008. He is presently an A.I.C.T.E. Emeritus Fellow at the CSE Department of J.U. His current fields of research interest include pattern recognition, image processing, and multimedia systems. He is a senior member of the IEEE, U.S.A., Fellow of I.E. (India) and W.B.A.S.T., Kolkata, India and a former Fellow, Alexander von Humboldt Foundation, Germany.

Variational Iteration Method For Solving Nonlinear Fractional Integro-Differential Equations

Muhammet Kurulay[1], Aydin SECER[2]

[1]Yildiz Technical University,
Faculty of Art and Sciences Department of Mathematics,
34210-Davutpasa-İstanbul, Turkey,
[2]Ataturk University,
Faculty of Art and Sciences Department of Mathematics,
25000-Erzurum, Turkey,
Corresponding Author Adress: aydinsecer@gmail.com

***Abstract*:** In this study, a modification of variational iteration method is applied to solve nonlinear fractional integro-differential equations. The fractional derivative is considered in the Caputo sense. The approximate solutions are calculated in the form of a convergent series with easily computable components. Through examples, we will see the modified method performs extremely effective in terms of efficiency and simplicity to solve nonlinear fractional integro-differential equations.

***Keywords*:** Nonlinear fractional integro-differential equations; Fractional derivative; Variational iteration method.

1. Introduction

In recent years, it has turned out that many phenomena in physics, engineering, chemistry, and other sciences can be described very successfully by models using mathematical tool from fractional calculus, such as, frequency dependent damping behavior of materials, diffusion processes, motion of a large thin plate in a Newtonian fluid creeping and relaxation functions for viscoelastic materials. etc. In addition to use of fractional differentiation for the mathematical modeling of real world physical problems has been widespread in recent years, e.g. the modeling of earthquake, the fluid dynamic traffic model with fractional derivatives, measurement of viscoelastic material properties, etc.

Most fractional differential equations do not have exact analytic solutions. There are only a few techniques for the solution of fractional integro-differential equations. Three of them are the Adomian decomposition method [1], the collocation method [2], and the fractional differential transform method [3]. The variational iteration method was proposed by he [4-10] and has found a wide application for the solution of linear and nonlinear differential equations, for example, linear fractional integro-differential equations[4], nonlinear wave equations [5], Fokker–Planck equation [6], Helmholtz equation [7], klein-Gordon equations [8], integro-differential equations [9], and space and time-fractional KdV equation [10].

In the study presented, fractional differentiation and integration are understood in Caputo sense because of its applicability to real world physical problems. We will set a new modified variational iteration method to solve nonlinear fractional integro-differential equations. It will show the modification of the method is a useful and simplify tool to solve nonlinear fractional integro-differential equations as used in other fields.

2. Basic Definitions

In this section, let us recall essentials of fractional calculus first. The fractional calculus is a name for the theory of integrals and derivatives of arbitrary order, which unifies and generalizes the notions of integer-order differentiation and n-fold integration. We have well known definitions of a fractional derivative of order $\alpha > 0$ such as Riemann–Liouville, Grunwald–Letnikow, Caputo and Generalized Functions Approach [11,12]. The most commonly used definitions are the Riemann–Liouville and Caputo. For the purpose of this paper the Caputo's definition of fractional differentiation will be used, taking the advantage of Caputo's approach that the initial conditions for fractional differential equations with Caputo's derivatives take on the traditional form as for integer-order differential equations. We give some basic definitions and properties of the fractional calculus theory which were used through paper.

Definition 2.1. A real function $f(x), x > 0,$ is said to be in the space $C_\mu, \mu \in R$ if there exists a real number $(p > \mu)$, such that $f(x) = x^p f_1(x),$ where $f_1(x) \in C[0,\infty),$ and it said to be in the space C_μ^m iff $f^m \in C_\mu, m \in N.$

Definition 2.2. The Riemann–Liouville fractional integral operator of order $\alpha \geq 0,$ of a function $f \in C_\mu, \mu \geq -1,$ is defined as

$$J_0^v f(x) = \frac{1}{\Gamma(v)} \int_0^x \left(x-t\right)^{v-1} f(t)dt, \quad v > 0,$$

$J^0 f(x) = f(x).$

It has the following properties:

For $f \in C_\mu, \mu \ge -1, \alpha, \beta \ge 0$ and $\gamma > 1$:

$$1. J^\alpha J^\beta f(x) = J^{\alpha+\beta} f(x),$$

$$2. J^\alpha J^\beta f(x) = J^\beta J^\alpha f(x),$$

$$3. J^\alpha x^\gamma = \frac{\Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+1)} x^{\alpha+\gamma}.$$

The Riemann–Liouville fractional derivative is mostly used by mathematicians but this approach is not suitable for the physical problems of the real world since it requires the definition of fractional order initial conditions, which have no physically meaningful explanation yet. Caputo introduced an alternative definition, which has the advantage of defining integer order initial conditions for fractional order differential equations.

Definition 2.3. The fractional derivative of $f(x)$ in the Caputo sense is defined as

$$D_*^v f(x) = J_a^{m-v} D^m f(x) = \frac{1}{\Gamma(m-v)} \int_0^x (x-t)^{m-v-1} f^{(m)}(t) dt,$$

for $m-1 < v < m,\ m \in N,\ x > 0,\ f \in C_{-1}^m.$

Lemma2.1. If $m-1 < \alpha < m,\ m \in N$ and $f \in C_\mu^m, \mu \ge -1,$ then

$$D_*^\alpha J^\alpha f(x) = f(x),$$

$$J^\alpha D_*^v f(x) = f(x) - \sum_{k=0}^{m-1} f^k(0^+) \frac{x^k}{k!},\ \text{x>0}.$$

The Caputo fractional derivative is considered here because it allows traditional initial and boundary conditions to be included in the formulation of the problem.

Definition 2.4. For m to be the smallest integer that exceeds α , the Caputo time-fractional derivative operator of order $\alpha > 0$ is defined as

$$D_{*t}^\alpha u(x,t) = \frac{\partial^\alpha u(x,t)}{\partial t^\alpha} = \begin{cases} \dfrac{1}{\Gamma(m-\alpha)} \displaystyle\int_0^t (t-\xi)^{m-\alpha-1} \frac{\partial^m u(x,\xi)}{\partial \xi^m} d\xi, & \text{for } m-1 < \alpha < m, \\ \dfrac{\partial^m u(x,t)}{\partial t^m}, & \text{for } \alpha = m \in N \end{cases}$$

and the space-fractional derivative operator of order $\beta > 0$ is defined as

$$D_{*x}^\alpha u(x,t) = \frac{\partial^\beta u(x,t)}{\partial x^\beta} = \begin{cases} \dfrac{1}{\Gamma(m-\beta)} \displaystyle\int_0^x (x-\theta)^{m-\beta-1} \frac{\partial^m u(\theta,t)}{\partial \theta^m} d\theta, & \text{for } m-1 < \beta < m, \\ \dfrac{\partial^m u(x,t)}{\partial x^m}, & \text{for } \beta = m \in N. \end{cases}$$

3. Modification of the Variational Iteration Method

Concerning the general fractional integro-differential equation of the type

$$D^\alpha y(t) = f\left(t, y(t), \int_0^t k(s,y) ds\right) \quad (1)$$

where D^α is the derivative of $y(t)$ in the sense of Caputo, and $n-1 < \alpha < n\ (n \in N),$ subject to the initial condition

$$y(0) = c.$$

According to the variational iteration method (VIM), we can construct the following correction functional

$$y_{n+1}(t) = y_n(t) + I^\alpha F(t) \quad (2)$$

where $F(t) = \lambda\left[D_*^\alpha y_n(t) - f\left(t, y_n, \int_0^t k(s,y_n) ds\right) dt\right],\ y_n(t)$ is the nth approximation, and I^α

is Riemann-Liouville`s fractional integrate.

The lagrenge multiplier can not easy identified through (2), so approximation of the corrrection functional can be expressed as follows

$$y_{n+1}(t) = y_n(t) + \int_0^t \lambda\left\{\frac{d^n y_n(t)}{dt^n} - f\left(t, y_n(t), \int_0^t k(s,y_n) ds\right)\right\} dt. \quad (3)$$

Then the Lagrange multiplier can be easily determined by the variational theory in (3).

λ is a general Lagrange multiplier [13]. Lagrange multipliers

$$\lambda = 1, \quad \text{for} \quad n = 1.$$

Substituting the identified Lagrange multiplier into (2) result in the following iteration procedures

$$y_{n+1}(t) = y_n(t) - I^\alpha\left\{D^\alpha y_n(t) - f\left(t, y_n(t), \int_0^t k(s,y_n) ds\right)\right\}, (n = 0,1,2,...).$$

4. Numerical Experiment

In this section, we apply VIM to solve a nonlinear fractional integro-differential equations. All the results are calculated by using the symbolic computation software Maple.

Example

Consider the following system of nonlinear fractional integral–differential equations, with initial values

$$(n_1)_0(0)=N_1,\quad (n_2)_0(0)=N_2\,.$$

$$D_*^{\alpha}n_1(t)=n_1\left(K_1-\gamma_1 n_2-\int_{t-T_0}^{t} f(t-s)n_2(s)ds\right),\quad K_1>0,\ 0<\alpha\le 1,$$

$$D_*^{\alpha}n_2(t)=n_2\left(-K_2-\gamma_2 n_1-\int_{t-T_0}^{t} f(t-s)n_1(s)ds\right),\quad K_2>0.$$

To solve this system by VIM, let us consider;

$$(n_1)_{n+1}(t)=(n_1)_n(t)+I^{\alpha}\left\{\lambda\left[D_*^{\alpha}(n_1)_n(t)-g\left[(n_1)_n(t)\right]\right]\right\},$$

$$(n_2)_{n+1}(t)=(n_2)_n(t)+I^{\alpha}\left\{\lambda\left[D_*^{\alpha}(n_2)_n(t)-g\left[(n_2)_n(t)\right]\right]\right\},$$

where

$$g\left[(n_1)_n(t)\right]=n_1\left(K_1-\gamma_1 n_2-\int_{t-T_0}^{t} f(t-s)n_2(s)ds\right),$$

$$g\left[(n_2)_n(t)\right]=n_2\left(K_2-\gamma_2 n_1-\int_{t-T_0}^{t} f(t-s)n_1(s)ds\right),$$

$(n_1)_n(t)$ and $(n_2)_n(t)$ are n th approximation.
We start with

$$(n_1)_0(0)=N_1,\ (n_2)_0(0)=N_2,$$

by the variational iteration formula, we have

$$n_1(t)=N_1+\frac{N_1\left[K_1-\gamma_1 N_2-N_2(1-e^{-T_0})\right]t^{\alpha}}{\Gamma(\alpha+1)}+\frac{N_1K_1^2t^{2\alpha}}{\Gamma(2\alpha+1)},$$
$$n_2(t)=N_2+\frac{N_2\left[K_2-\gamma_2 N_1-N_1(1-e^{-T_0})\right]t^{\alpha}}{\Gamma(\alpha+1)}+\frac{N_2K_2^2t^{2\alpha}}{\Gamma(2\alpha+1)}.$$

When $\alpha=1$, then we have

$$n_1(t)=N_1+N_1\left[K_1-\gamma_1 N_2-N_2(1-e^{-T_0})\right]t+0.5N_1K_1^2t^2,$$
$$n_2(t)=N_2+N_2\left[K_2-\gamma_2 N_1-N_1(1-e^{-T_0})\right]t+0.5N_2K_2^2t^2,$$

which is the same solution given by Biazar [14].

5. Conclusion

In this paper, we applied the modified variational iteration method for solving the nonlinear fractional integro-differential equations. Comparison with other traditional methods, the simplicity of the method and the obtained exact results show that the modified variational iteration method is a powerful mathematical tool for solving nonlinear fractional integro-differential equations. The method was used in a direct way without using linearization, perturbation or restrictive assumptions. It may be concluded that the method is very powerful and efficient in finding analytical as well as numericalsolutions for wide classes of nonlinear fractional integro-differential equations. It provides more realistic series solutions that converge very rapidly in real physical problems.

6. References

[1] S. Momani, and A. Qaralleh, An efficient method for solving systems of fractional integro-differential equations, Comput. Math. Appl. 52,459–470, 2006.

[2] Rawashdeh EA, Numerical solution of fractional integro-differential equations by collocation method, Appl. Math. Comput. 176, 1–6, 2006.

[3] A. Arikoglu, and I. Ozkol , Solution of fractional integro-differential equations by using differential transform method, Chaos Soliton Fractals, 40, 521-529, 2007.

[4] Wen-Hua Wang, An effective method for solving fractional integro-differential equations, Acta universitatis apulensis, 20, 2009.

[5] J. -H. He, Variational iteration method a kind of non-linear analytical technique, J. Nonlinear Mech. 34, 699–708, 1999.

[6] M. Dehghan, and M. Tatari, The use of He`s variational iteration method for solving the Fokker–Planck equation, Phys. Scripta. 74 , 310–316, 2006.

[7] S. Momani, and S. Abuasad, Application of He's variational iteration method to Helmholtz equation , Chaos Soliton Fractals. 27, 1119–1123, 2006.

[8] S. Abbasbandy, Numerical solution of non-linear Klein-Gordon equations by variational iteration method, Internat. J. Numer. Methods Engrg. 70, 876–881, 2007.

[9] S. -Q. Wang, and J.-H. He, Variational iteration method for solving integro-differential equtionas, Phys Lett. A. 367, 188–191, 2007.

[10] S. Momani, Z. Odibat and A. Alawneh, Variational iteration method for solving the spaceand time-fractional KdV equation, Numer. Methods Partial Differential Equations. 24(1), 262–271, 2008.

[11] I. Podlubny Fractional differential equations. San Diego: Academic Press; 1999.

[12] M. Caputo, Linear models of dissipation whose Q is almost frequency independent. Part II, J. Roy. Austral. Soc. 13, 529–539, 1967.

[13] Inokuti M, Sekine H, Mura T. General use of the Lagrange multiplier in non-linear mathematical physics. In: S. Nemat-Nasser (editor). Variational method in the mechanics of solids. Oxford: Pergamon Press; p. 156–62, 1978.

[14] J. Biazar, Solution of systems of integral-differential equations by Adomian decomposition method, Applied Mathematics and Computation 168 (2), 1232-1238, 2005.

Phishing E-mail Analysis

Shamal Firake, Pravin Soni, Dr. B.B.Meshram

shamal@inbox.com pravindsoni@gmail.com bmeshram@vjti.org.in

Abstract— The act of sending a forged e-mail (using a bulk mailer) to a recipient, pretending to be a legitimate in order to scam the recipient into divulging private information such as credit card numbers or bank account passwords is known as phishing. Seeking sensitive user data is the primary objective of the phishing e-mails. With the increase in the online trading activities, there has been a phenomenal increase in the phishing scams which have now started achieving monstrous proportions. According to Gartner estimates, 3.3% of the 124 million consumers who received phishing email last year were victimized and lost money because of the phishing email attacks. This paper is centered around Phishing Attacks on E-mail. Paper contains the brief literature review about the different approaches developed to detect and prevent phishing attacks. Paper gives basics of e-mail phishing attack, as how to send forged e-mails and how one can send mass emails to someone. We have also given the method for e-mail traversing which will be used for e-mail forensic analysis of forged e-mails. We have also given our proposed system for Detection and Prevention of Phishing Attacks on E-mail.

Keywords— *E-mail , Phishing Attack, E-mail Forging, Mass E-mailing, HyperLink Detection , Digital Signature.*

I. Introduction

Phishing has actually been around for over 15 years, starting with America Online (AOL) back in 1995.There were programs (like AOHell) that automated the process of phishing for accounts and credit card information.[2] Actually the term *phishing* is derived from the fact that Internet scammers "fish" for users' financial information and password data. "Ph" is a common replacement for the letter "f" in hacker lingo; one of the earliest forms of hacking was known as "phone phreaking."Phishing, in computer security field is described as the criminally fraudulent process that attempts to acquire sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity in an electronic communication. The frequently used attack method is to send e-mails to potential victims, which seemed to be sent by banks, online organizations, or ISPs. In these e-mails, they will makeup some causes, e.g. the password of your credit card had been mis-entered for many times, or they are providing upgrading services, to allure you visit their Web site to conform or modify your account number and password through the hyperlink provided in the e-mail. You will then be linked to a counterfeited Web site after clicking those links.

The phishing problem is a hard problem for a number of reasons. Most difficulties stem from the fact that it is very easy for an attacker to send mass emails or to do email forging to spoof e-mail addresses. Previous work indicates that the ability to create exactly similar looking copies of legitimate e-mails, as well as users' unfamiliarity with browser security indicators, leads to a significant percentage of users being unable to recognize a phishing attack. Unfortunately, the ease with which copies can be made in the digital world also makes it difficult for computers to recognize phishing attacks. As the phishing websites and phishing emails are often nearly identical to legitimate websites and emails, current filters have limited success in detecting these attacks, leaving users vulnerable to a growing threat.

According to the statistics provided by the Anti-Phishing Working Group (APWG) [1], in March 2010, email reports received by APWG from consumers were 30,577.The number of unique phishing sites detected ,in March 2010 were 29879.Payment Services returned to being the most targeted industry sector after Financial Services held top position during H2 2009. However, the category of ‚Other' rose from 13 percent to nearly 18 percent from Q4 2009 to Q1 2010, an increase of nearly 38 percent. Amongst the affected industry sector Payment services hold 37% and Financial services 35.9%[20].

In this paper we study the phishing e-mail analysis. We will see the basic steps how e-mail phishing attack is carried out. The easiest method to disguise the source of an e-mail and send it to the victim pretending to be a legitimate one is , E-mail Forging. Most attackers use this technique to fool the victim into believing that somebody else has sent the particular e-mail [34]. The basic experimentation of e-mail forging is also discussed. Another technique ,which is carried out e-mail phishing attack is Mass E-mailing. Now a days number of free softwares are there like MassMailer, BulkMail , eMailer, Mail me 3.00 etc. The only problem with these tools is that , they require to be on client side and require attacker to be stay online. To overcome this problem the Romanian Phishers found a lazy bulk mailing tool written in PHP [2]. This PHP bulk-mailing tool that executes on the server side, which utilizes the bandwidth of the compromised dedicated server. The PHP scripts used and their experimentation is included here. Thus by using these all techniques it is very easy to carry out e-mail phishing attacks that are the major threat in today's world. To overcome this problem till now many solutions are proposed which are really useful in their own sense. The brief discussion of all these approaches is done in literature survey .We also propose our own model to detect as well as prevent e-mail phishing attacks.

The rest of the paper is organised as follows. In section II the brief literature survey is provided which includes E-mail basics and approaches used to detect or prevent phishing attacks. In e-mail basics we have discussed e-mail header, mail delivery process and anonymous e-mails. Section III gives experimentation and analysis of email phishing attack. In section IV the proposed model to detect and prevent phishing attack is given. Section V concludes the paper.

II. LITERATURE SURVEY

A. E-mail Basics

E-mail contains specific key elements that enable it to communicate and route to the correct places. The design of the e-mail system is what makes e-mail one of the most efficient forms of communication today. Ironically, the e-mail system's infrastructure is similar to that of the traditional post office in that it requires you to have "routable" addresses enabling mail to be delivered. The mail server is similar to your human mail carrier, and the mail client is you physically walking to your mailbox.

1]. E-mail Header

Each time an email is sent on the Internet, it not only carries the message body , but also transmits relevant information on the path travel by it.This information is known as Email Header of the email. The most effective and easiest way to trace an email is to analyze its email headers. Most cybercrime investigators turn to email headers for evidence in any kind of e-mail related crime. All email communications on the internet is governed by rules and regulations laid down by two different protocols:

- Simple Mail Transfer Protocol(SMTP port 25)
- Post Office Protocol(POP port 110).

Each email on the internet originates at the sender's post office server with the help of SMTP commands.It is routed via number of intermediate mail servers and then finally reaches to the destination post office where the receiver use POP commands to download it to local system.Email headers are automatically generated and embedded into an email message both during composition and transfer between systems. They not only contain valuable information on the source of the email , but also represent the exact path taken by it, which can be represented as

Sender Outbox Source → **Mail Server** → **Intermediate Mail Server** → **Destination Mail Server**

→ **Destination Inbox**.

A typical email header looks something like this:

```
Return Path:
pankaj.chandigarh@gmail.com
Received: from
pankaj.chandigarh@gmail.com by
(208.50.6.127:25) via
   ug-out-1314.google.com
(66.249.92.171:20204) with [InBox.Com
SMTP Server]id 608240002387.WM27 for
shamal@inbox.com; Thu, 24 Aug 2006
00:34:37-0800
Received: by ug-out-1314.google.com
with SMTP id y2so389480uge        for
<shamal@inbox.com>; Thu, 24 Aug 2006
01:34:17 -0700 (PDT)
DomainKey-Signature: a=rsa-sha1; q=dns;
c=nofws;s=beta;d=gmail.com;
h=received:message-
id:date:from:to:subject:mime-
version:content-type;
b=Gq4HzLHTQwXmfvsJcJ65quwYgG9l/a6zLWwPB
r63PZ1WJE/8thjWVm+BD8GWCCg6Iu6/CdHYGggV
pcpXqmfNO4JDtVulPMkirNemUaSltKzjfuHF2DI
Ji1Zorhvq5CvxT10gTl92UjmQE5XMZkpHqGUBSm
X6O7Qwd27kfpZHjXo=
Received: by 10.67.119.13 with SMTP id
w13mr766642ugm;Thu, 24 Aug 2006
01:34:14 -0700 (PDT)
Received: by 10.66.237.1 with HTTP;
Thu, 24 Aug 2006 01:34:12 -0700 (PDT)
 Message-ID:
<1c058b0f0608240134s60d85438m67d196073f
8e1f14@mail.gmail.com>
 Date: Thu, 24 Aug 2006 14:04:12 +0530
 From: "Pankaj Mishra"
<pankaj.chandigarh@gmail.com>
 To: shamal@inbox.com,
shamal.firake@gmail.com
 Subject: Bjarne Stroustrup Book (C++)
 MIME-Version: 1.0
 Content-Type:multipart/mixed;
        boundary="----
=_Part_156068_33357618.1156408452495"
 X-Spam-Ratio: 0.02------
=_Part_156068_33357618.1156408452495
 Content-Type: multipart/alternative;
        boundary="----
=_Part_156069_14038231.1156408452495"

 ------
```

One can possibly identify the source of the email by simply Reverse Engineering the path travelled by it , which is explained in the next section.

2]. Mail Delivery Process

All e-mail headers contain the server and client information that controls the process of mail delivery. Many people who use e-mail clients have probably heard of SMTP servers and POP3 servers. Within the typical setup for e-mail, two ports are typically used: port 25, and port 110. Port 25 is the Simple Mail Transfer Protocol (SMTP), and its job is to transmit and receive mail—basically what is called a Mail Transfer Agent, or MTA.

An MTA is comparable to the mail carrier who picks up the mail and sends it off to where it needs to go. Just as the mail carrier drops off and picks up mail, so does the MTA. Port 110 is the Post Office Protocol, version 3 (POP3), and it is essentially the mailbox from which users pick up their mail up. The mail server infrastructure works in such an efficient fashion that we did not use only four servers but, at minimum, eight servers to deliver our e-mail. In the process of sending e-mail, we query multiple DNS servers to obtain information about where the mail servers are on the Internet. Here is an example of the complete process for sending an e-mail

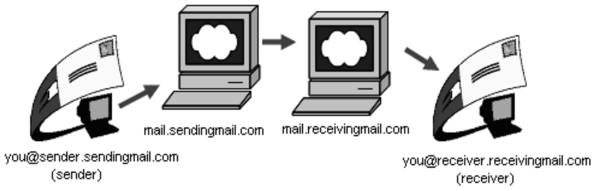


Figure 1 Standard Email Process

3]. Anonymous E-mail

Technology sector experts well know that SMTP was not designed with security in mind. E-mail is trivial to forge, and in more than one way, forged e-mail can be passed with ease to the mail transport agent (SMTP server). As we already are aware, spammers forge e-mails, and since phishers are classified as spammers, they take on this practice as well. Most spammers tend to forge e-mails for anonymity, since they are sending you annoying e-mails that will usually get a negative reaction, and if the e-mails were easily traceable, they would probably be caught.

Phishers forge for a different reason:They are attempting to con you, and they are using forgery to spoof a likely bank e-mail, such as verify@citibank.com. Not all headers can be forged, so the good news is that you can still track down the originator IP address, but unfortunately the phishers are not e-mailing directly from their homes.

The headers that can be forged are:

■ *Subject, Date, Message-ID*

■ Recipients: *From,To, CC*

■ Content body

■ Any arbitrary headers such as the *X-Mailer* and *X-Message-Info*

■ The initial *Received* headers

The headers that cannot be forged are:

■ The final *Received* headers

■ The originating mail server, including:

■ IP address

■ Subsequent timestamps

General clues within the header usually identify whether it is forged or not. The obvious one is the *Received* headers being inconsistent with mismatched *From* and *by* fields. The *HELO* name does not match the IP address, there are nonstandard headers in general placed within the e-mail, and wrong or “different” formats of the *Date, Received, Message-ID,* and other header labels.

B. *Different Approaches Developed To Detect And Prevent Phishing Attacks*

PILFER *et al.* [3] developed the tool *which* can be either deployed in a standalone configuration without a spam filter to catch a large percentage of phishing emails. PILFER Phishing email filter combines a set of features aimed at catching deception along with advanced machine learning techniques. The features used in approach includes Age of linked-to domains ,number of domains linked to, presence of attention-directing links (“click here”) that link to a domain other than the most common one in the email.

Engin Kirda et. al. **[6]** have developed an anti phishing solution called AntiPhish to guard users against a spoofed web site based phishing attack. The tool keeps track of the sensitive information of a user and generates warnings whenever sensitive information is typed on a form generated by a website that is considered untrusted . One of the drawbacks of the solution is that it lets the user go up to a stage where he is allowed to type in sensitive information on a form and then if the tool finds out that the website is untrustworthy; it warns the user against it. The user is thus susceptible to losing his sensitive data if the phisher employs tools such as a key-logger or a malware which is programmed to send screenshots of the user’s console every few seconds.

Juan Chen et. al. **[4]** have proposed an algorithm named LinkGuard which analyzes the generic characteristics of the hyperlinks in the phishing emails to deduce whether a site is spoofed or not. The algorithm makes use of a set of rules to analyze the URL viz, mismatch between the actual destination link and the link as seen by the user, use of IP addresses in dotted decimal format, absence of destination information in the text as seen by the user, etc.

Kapil et.el. **[16]** have developed an end user application that makes use of user provided data to check the authenticity of the destination URL and hence is able to give a more accurate prediction about the validity of the destination website. The approach save a mapping between supplied credentials and corresponding trusted website domains during the learning phase. In a detection phase, a submitted credential is matched with the saved credentials, and the current domain name is compared with the saved domain names. If there is a mismatch, a website is suspected as phishing.

Kristofer Beck and Justin Zhan [9] proposed Solution named Thin Client. A thin client is created to allow a secure connection between a client and the institution. We believe that this is a better way to prevent people from losing their private information due to phishing. A different interface than the traditional browsers which prove through past research are prone to fail in complete securityThis algorithm in itself may have faults because the phisher can theoretically take time and reengineer our thin client. This would be a change in the way phishers usually spoof a website by using phishing kits to replicate HTML code. It is harder to reengineer ActionScript used to create the thin client.

Ben Adida et.el.[13] suggested the Trusted Email Approach proposed the solution to authenticate certain email messages for the purpose of distinguishing legitimate business emails from spam and phishing attempts. All of the problems with spam and phishing start with SMTP. Due to its un-authenticated nature, anybody can send an email with a *From field equal to, for example,* ”billingsupport@companyXYZ.com”.A number of attempts have been made to add authentication to email. Most notably, PGP and S/MIME provide tools for encrypting and signing of email messages. A recipient of a signed message can verify the original sender based on the cryptographic signature. Unfortunately, neither PGP nor S/MIME would work on such large scale. The solution uses public key certificates for institutions only (though it does not require certificates) and does not require that users obtain certificates or public-private key pairs themselves .Unlike other solutions, *Trusted Email does not require* modifications to the Internet infrastructure (e.g. SMTP, DNS, etc.).

The proposed method of email verification is not designed to provide protection over already compromised communications channels. Lacking a trusted central key repository means that the initial communication between the user and the third party must be made without cryptographic verification of the third party’s identity. The *Trusted Email system is vulnerable* to a man-in-the-middle attack. It is also vulnerable to an eavesdropping attack where the attacker is able to eavesdrop the custom message, create his own email containing the attacker’s public key and the custom message and send the message so that it arrives before the original bank’s message.

Gansterer Polz et.el.[15] done e-mail classification for phishing defense based on different features of an-email. It is a classification-based approach for filtering phishing messages in an e-mail stream. Various features of every e-mail are extracted. This forms the basis of a classification process which detects potentially harmful phishing messages. The approach introduces new sixteen features. These newly introduced features belong to three different groups:

- The first group contains six “off-line” features.
- The second group contains eight “online” features.
- The third group is a control group of presumably class independent features containing two features: Subject length (SubjectLen) counts the number of characters in the subject field, and Sender length (SenderLen) counts the number of characters in the sender field of a message.

Chandrashekharan Krishnan et.el.[7] analyzed the structural properties of e-mail to separate phishing mails from legitimate e-mails. The main goal of approach is to classify phishing emails using a set of characteristics that remain relatively invariant across a large amount of emails. The characteristics used are language, layout, structure of phishing email so that all different contexts of phishing emails can be captured . The features relevant to language, composition and writing such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage , stylistic and sub-stylistic features will remain relatively constant. Identifying and learning of these structural features with sufficiently high accuracy is very difficult challenge during phishing email classification.

III. EXPERIMENTATION AND PHISHING EMAIL ANALYSIS

Reading e-mails has become a dangerous activity. E-mails can carry dangerous viruses, worms which can be executed by merely opening e-mail or clicking on active link or picture in an e-mail. This e-mail phishing attacks are easily carried out by email spoofing. The two techniques known as e-mail forging and mass emailing made the task very easy for phishers to grab more number of victims. In this section we will see the basic steps of e-mail phishing attack and the two mostly used above mentioned techniques to carry out these e-mail phishing attacks. The last part of section gives the method to trace the source of an e-mail and a way to detect and trace the forged e-mail.

A. Steps of E-mail Phishing Attack

• The attacker obtains E-mail addresses for the intended victims. These could be guessed or obtained from a variety of sources.

• The attacker generates an E-mail that appears legitimate and requests the recipient to perform some action.

• The attacker sends the E-mail to the intended victims in a way that appears legitimate and obscures the true source.

• Depending on the content of the E-mail, the recipient opens a malicious attachment, completes a form, or visits a web site.

• The attacker harvests the victim's sensitive information and may exploit it in the future.

Thus attacker obtains the e-mail addresses of victims from internet and address list that user believed to be private(CNET).To send these phishing e-mails to victims attackers may use E-mail Forging or Mass emailing which are very easy to implement.

B. E-mail Forging

E-mail forging allows an attacker to disguise the source of an e-mail and send it to the victim [34]. Attackers use this technique to fool the victim into believing that mail has came from some legitimate source. Unfortunately ,there is very little that a victim can do to counter e-mail forging other than remain cautious and alert.

The Simple Mail Transfer protocol (SMTP) is the de facto standard protocol used by e-mail clients and daemons to send e-mails on the Internet. This protocol is used by the SMTP daemon that by default runs on Port 25 of a mail server. Each time user writes an e-mail and clicks on the SEND button , the e-mail client automatically issues SMTP commands to the remote mail server and sends the specified message.

Unfortunately , the SMTP protocol also makes it extremely easy for an attacker to send forged e-mails to a remote user.It is quite possible for a user to connect manually to the SMTP port(25) of a remote mail server and use SMTP commands to manually send an artificial e-mail. This process of using SMTP commands to send e-mails from someone else's e-mail account is known as ***E-mail Forging.***

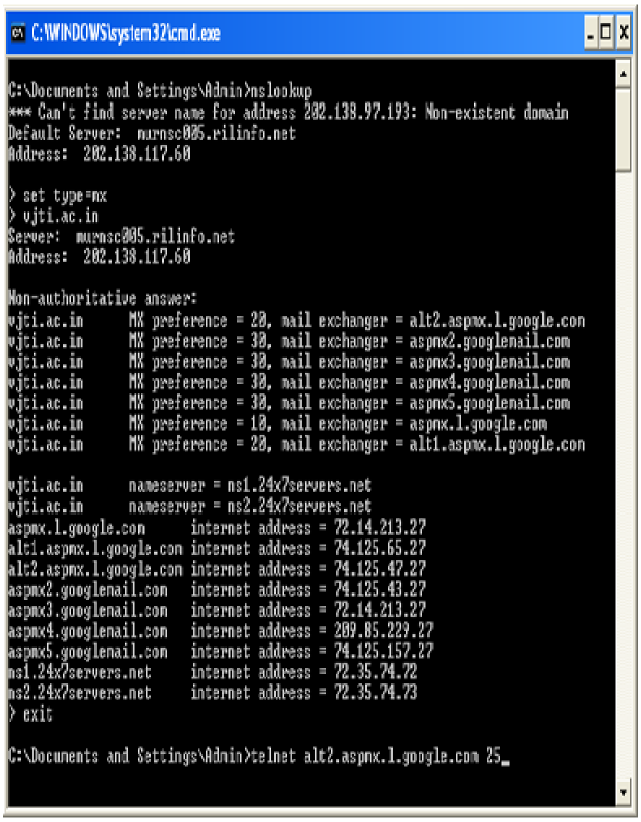
1]. The basic steps to carry out e-mail forging are:

- ***Step 1:*** **Open a command prompt**. In Windows, you can do this by clicking <Start>, <Run>, and type <cmd> in the box and press <OK>. You should get a black "Command Prompt" screen.
- ***Step 2:*** **You will need an SMTP server address to proceed**. Here is how to find one: On the command-line, type **<nslookup>**.
  - Then, type <set type=mx>.
  - Finally, enter the name of any website, for instance, <hazemdesigns.srhost.info>.
  - This will return the following: <Non-authoritative answer:hazemdesigns.srhost.info mail exchanger = 0 ASPMX.L.GOOGLE.COM.>.
  - The <aspmx.l.google.com> part is what you need. This is an SMTP server address.

  Type <exit> to exit out of nslookup. Fig 2 explains step 2.
- ***Step 3:*** Now after u have find out the smtp mail server type in cmd ***telnet "mail server" 25***

In this case mail server is ***alt2.gmail-smtp-in.l.google.com***

- ***Step 4:*** Now we will be connected to the mail server, so now to begin, type in ehlo for esmtp (extended smtp) and helo for smtp type server in command window that appear after last step.



```
C:\Documents and Settings\Admin>nslookup
*** Can't find server name for address 202.138.97.193: Non-existent domain
Default Server:  murnsc005.rilinfo.net
Address:  202.138.117.60

> set type=mx
> vjti.ac.in
Server:  murnsc005.rilinfo.net
Address:  202.138.117.60

Non-authoritative answer:
vjti.ac.in      MX preference = 20, mail exchanger = alt2.aspmx.l.google.com
vjti.ac.in      MX preference = 30, mail exchanger = aspmx2.googlemail.com
vjti.ac.in      MX preference = 30, mail exchanger = aspmx3.googlemail.com
vjti.ac.in      MX preference = 30, mail exchanger = aspmx4.googlemail.com
vjti.ac.in      MX preference = 30, mail exchanger = aspmx5.googlemail.com
vjti.ac.in      MX preference = 10, mail exchanger = aspmx.l.google.com
vjti.ac.in      MX preference = 20, mail exchanger = alt1.aspmx.l.google.com

vjti.ac.in      nameserver = ns1.24x7servers.net
vjti.ac.in      nameserver = ns2.24x7servers.net
aspmx.l.google.com      internet address = 72.14.213.27
alt1.aspmx.l.google.com internet address = 74.125.65.27
alt2.aspmx.l.google.com internet address = 74.125.47.27
aspmx2.googlemail.com   internet address = 74.125.43.27
aspmx3.googlemail.com   internet address = 72.14.213.27
aspmx4.googlemail.com   internet address = 209.85.229.27
aspmx5.googlemail.com   internet address = 74.125.157.27
ns1.24x7servers.net     internet address = 72.35.74.72
ns2.24x7servers.net     internet address = 72.35.74.73
> exit

C:\Documents and Settings\Admin>telnet alt2.aspmx.l.google.com 25_
```

Figure 2. Using nslookup utility in Windows to get mail server address

- ***Step 5:*** Now here type exactly as given below :
  - Type:-helo mail server : the name of mail server
  - Type:- mail from: email id from where this mail is send from.eg. mail from:
  - Type:- rcpt to: email id to whom this mail go to.
    eg. rcpt to:
  - Now type:- data
  - Type the message that you want to send.
  - Now at last type .(dot) after entering data.

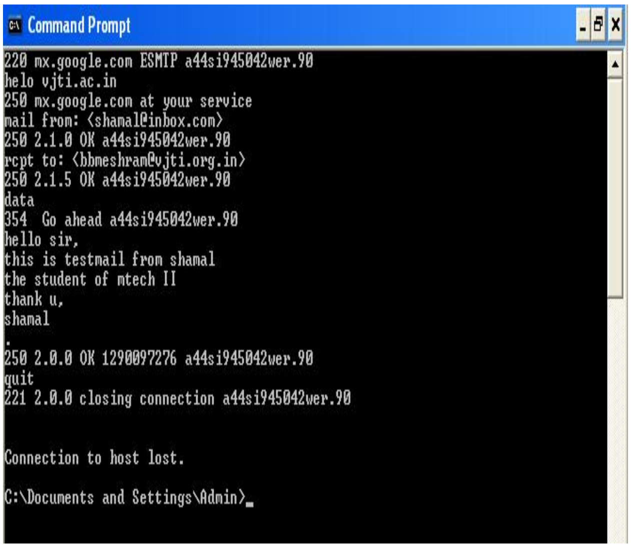
The step 4 and 5 are explained in fig 3.

In reality , one does not need to remember any SMTP commands. You can get help by simply typing HELP. To get the details of specific command type HELO followed by command name.

2]. Advanced E-mail forging

In the last section, we have seen the basic e-mail forging which is very easy to execute. However ,an attacker requires more control over the various features of the forged e-mail. Thus advanced features of e-mail forging includes:The Subject Field

- Sending File attachments
- The CC & BCC Fields
  - Using rcpt to Command
  - Using CC Field



```
220 mx.google.com ESMTP a44si945042wer.90
helo vjti.ac.in
250 mx.google.com at your service
mail from: <shamal@inbox.com>
250 2.1.0 OK a44si945042wer.90
rcpt to: <bbmeshram@vjti.org.in>
250 2.1.5 OK a44si945042wer.90
data
354  Go ahead a44si945042wer.90
hello sir,
this is testmail from shamal
the student of mtech II
thank u,
shamal
.
250 2.0.0 OK 1290097276 a44si945042wer.90
quit
221 2.0.0 closing connection a44si945042wer.90


Connection to host lost.

C:\Documents and Settings\Admin>_
```

Figure 3 Using SMTP commands to send fake e-mail

- **Using Subject Field**

Most professional and personal e-mails on the Internet have a suitable subject field describing the contents of the e-mail. Hence, from an attacker's perspective , in order to reduce suspicion , it is extremely important to send a forged e-mail with a subject .The SUBJECT argument is accepted by DATA command that is normally used to specify the content of forged e-mail. As soon as an attacker enters the DATA command the SMTP prompt is ready to accept both the contents of e-mail and also arguments if any. Fig 4 explains how to use subject field.

- **Sending File attachments**

All e-mail attachments were transmitted across networks using Unix-to-Unix encoding standard(UU-encoding standard).

UU-encoding Standard

- Transmit data safely without any corruption or loss of bytes.
- Converts data files into ASCII format
- Increases the size of any file by 42%

Thus files can be attached to forged e-mail by following the steps:

- Converting the file to be attached into the uuencoded format

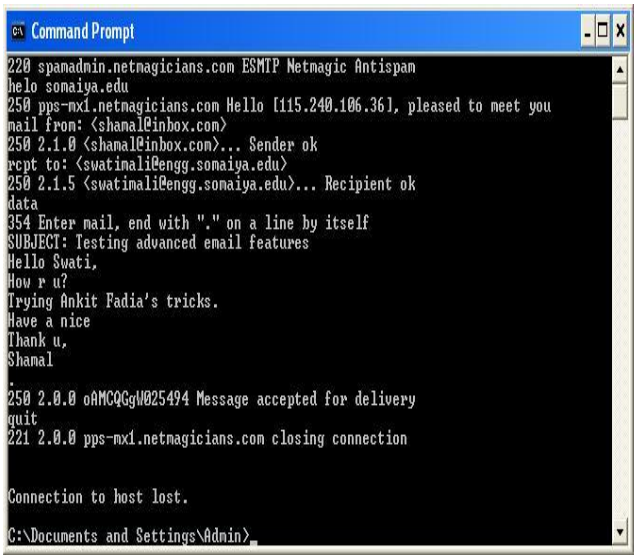
Connecting to the remote mail server and pasting the uuencode obtained in step1 into the DATA command.

Fig 5. explains how to send file attachments in e-mail forging.

- **Multiple Entries in TO Field**

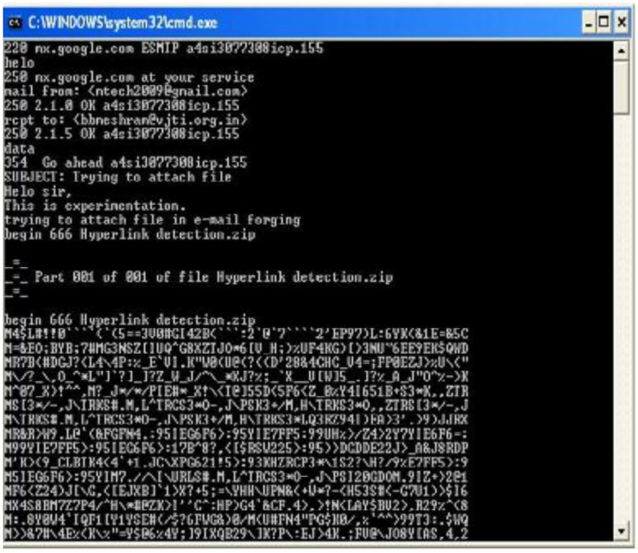
1. Connect and exchange introductions with mail server.
2. Use multiple RCPT commands to send the same e-mail to more than one persons.

The fig 6's telnet session demonstrates how to enter multiple e-mail addresses in the TO field



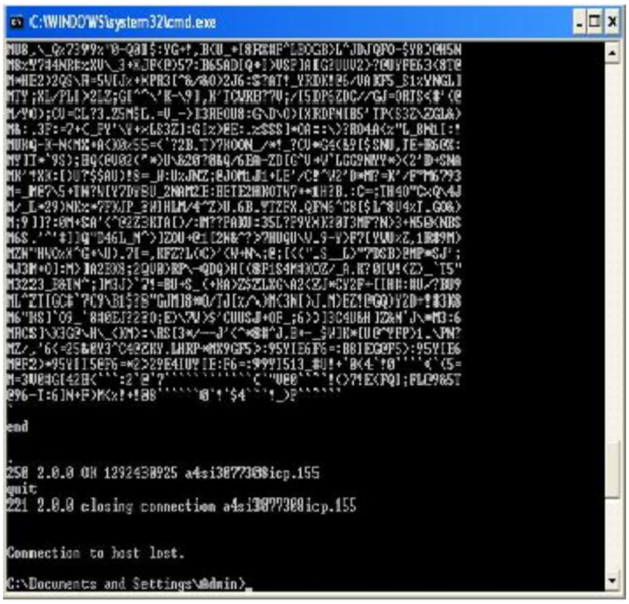
```
220 spamadmin.netmagicians.com ESMTP Netmagic Antispam
helo somaiya.edu
250 pps-mx1.netmagicians.com Hello [115.240.106.36], pleased to meet you
mail from: <shamal@inbox.com>
250 2.1.0 <shamal@inbox.com>... Sender ok
rcpt to: <swatimali@engg.somaiya.edu>
250 2.1.5 <swatimali@engg.somaiya.edu>... Recipient ok
data
354 Enter mail, end with "." on a line by itself
SUBJECT: Testing advanced email features
Hello Swati,
How r u?
Trying Ankit Fadia's tricks.
Have a nice
Thank u,
Shamal
.
250 2.0.0 oAMCQGgW025494 Message accepted for delivery
quit
221 2.0.0 pps-mx1.netmagicians.com closing connection


Connection to host lost.

C:\Documents and Settings\Admin>_
```

Figure 4 Advanced e-mail forging using subject field



```
220 mx.google.com ESMTP a4si3077308icp.155
helo
250 mx.google.com at your service
mail from: <ntech2009@gmail.com>
250 2.1.0 OK a4si3077308icp.155
rcpt to: <bbmeshram@vjti.org.in>
250 2.1.5 OK a4si3077308icp.155
data
354  Go ahead a4si3077308icp.155
SUBJECT: Trying to attach file
Helo sir,
This is experimentation.
trying to attach file in e-mail forging
begin 666 Hyperlink detection.zip

_=_
_=_ Part 001 of 001 of file Hyperlink detection.zip
_=_

begin 666 Hyperlink detection.zip
M4$L#!!0````(`(5==3V0#GI42B(```:2`@'7````2'EP97)L:6YK<&1E=&5C
M=&EO;BYB;7#MG3NSZ[1UQ^GBXZTJO*6[V_H;)%UF4KG)[)3NU"6EE9EK$QWD
MR7B<#DGJ?<L4\4P:%_E`V1.K"W0<U@<?<<D'28&4CHC_V4=;FP0EZJ>%U\<"
M\/?_\,O_^*L"]`?]_]?Z_W_J/^\_*KJ?%;_`X__U[WJ5_.]?%_A_J"O^%->K
M^07_X)!^^,M?_J*/*/PIE#*_X!\<I@J55D<5F6<Z_0%Y4I651B+S3*K,,ZTR
MS[3*/-,J\IRKS#.M,L^TRCS3*O-,J\PSK3+/M,H\TRKS3*O,,ZTRS[3*/-,J
M\TRKS#.M,L^TRCS3*O-,J\PSK3+/M,H\TRKS3*LQ3RZ94I)EA>3'.>9)JJRX
MR&R>W9.L@'(&FGFN4.:95IEG6F6>:95YIE7FF5:99UH%>/Z4>2Y7YIE6F6=:
M99YIE7FF5>:95IEG6F6>:17B^8?,<I$RSV225>:95))DCDDE22J>_A&J8RDP
M'K)<9_CLBIK4<4`+1.JC\XPG621!5):93KHZRCP3*\1S2?\H?/9%E7FF5>:9
M5IEG6F6>:95YIM7./^I\URLS#.M,L^TRCS3*O-,J\PSI2@GDOM.9IZ+>2@1
MF6<Z24)J[\G,<[EJXB]`1>X?+5;=\YHH\UPN&<+U*?-<H53S#<-G7U1>>$I6
MX4S8BM7Z7P4/^H\*#@ZK>1''C^:HP>G4`&CF.4>.>!N<LAY$BU2>.R29%^<8
M:.8Y0V4'IQF1[Y1YSE#</$?6FVG&)@/M<U#FN4"PG$K@/,%`^^>99T3:.$WQ
M>>&7#\4E%<K\%"=Y$@6%4Y;]9IXQB29\]K?P\:EJ>4K.;FU@\JO8Y[AS,4,2
```

Figure 5 Advanced e-mail forging to attach files(Part A)

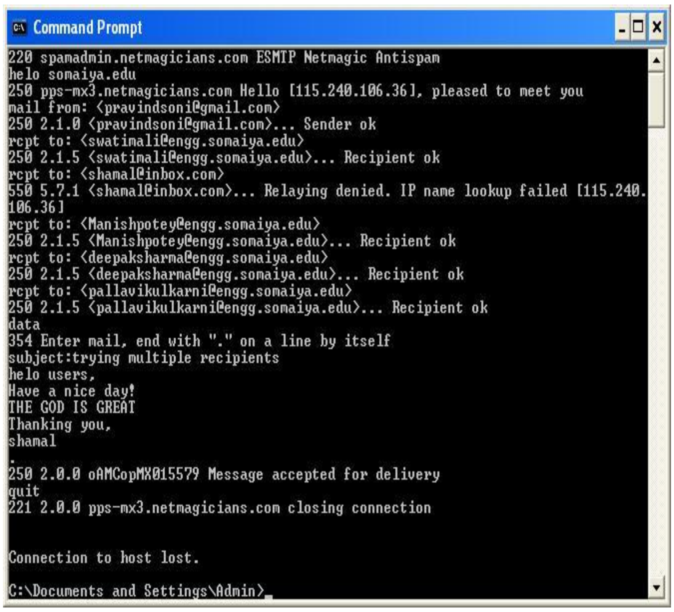


```
MU8_.\_Q&73?9%'0-Q0I$:YG+!,B(U_+I8R&HF^LDDGB>L^JDJQFO-$Y8)CH5N
M8&Y7I4NRE&XU\_3+&JF(@)57:B65ADIQ+I>USP]AIG2UUV2>?C0YFE63<8T@
M+HE2)2QS\H=5V[J&+KPR3[^%/&0)2J6:S?AT!_YRDK!26/VAKF5_S1&YNGL]
MTY;XL/PLI>2LZ;GI^^\'K-\9],K'TCVRB??V;/I5DP6ZDC//GJ=ORIS<#'(@
M/Y0);CU=CL?3.Z5M$L.=U_->I3HEOU8:G\D\0>IXRDF4IB5'IP(S3Z\ZGL&)
M&:.3F:=7+C_PY'\Y+&LS3Z]:GI&)@E:.&SSS]+CA==\)?RO4A<&"L_BM1I:!
MUHQ-X-N(MX+A<)X&55=<'?2B.T)7HOON_/*!_?CU+G4(&9[$SNU,TE+B6@X:
MY]I*'9S);BQ(@V02("*)U\&20?8&Q/6EA-ZD[G^V+V'LGG9NYY*)<2'D+SNA
MK'!XX:[)U?$$AU)!8=_H:U&JNZ;@JOM1J1+LE'/C!^V2'D+M?=K'/F"M6?93
M=_ME?\5+IW?V[Y7DYBU_2NAM2E:BETE2HNOTN?++1H?B.:C=;IH40"C&Q\4J
M/_L+29)NK&+7FXJP_2H]HLM/4^Z>U.6B_YTZFX.QFN6^CB[$L^8U4&T_GO&)
M;9]]?:0M+SA'<^@2Z3KTA[)/:M??PAXU=3SL?F9YXK?0I3MF?N>3+N5B(NBS
M6S.'^'#]]Q"D46L_M^>]ZOU+@1[2H&^?>?HUQU\V_9-Y>F7[YUU&Z,1R#9M)
MZH"HVO&X^G+\U).7I=,RFZ?L(C>'<V+N\;@;[<("_S__L>"7DSB)@MP*SJ';
MJ3M+O]:M>]A2EX8;2QUB)RP\-QDQ>H[(@F1S4M#XOZ/_A.K?0IV!<Z>_`T5"
M3223_B&IN^;]M3J>`7!=BU+S_(+HA)ZSZLXG\A2<ZJ*CY2F-[IHH:HU/?BU9
L^ZI[GC#`?C9\B1$?B"GJM]8*0/TJ[X/^\)M(3N[)J.M)EZ!@GQ)Y2D+!#3]X8
M6"HS]`O9_`8H0EJ?2=0;E)\7U>S'CUUSJ+OF_;6>)]3C4U&H]Z&N'J\*M3:6
MACS]\X3G@\H\_<XM>=\RS[3*/--J'<^*@#^J.B+-_$VJK*[U@^YFP)1.\PN?
MZ/,'6<=25&@Y3^C4@ZKY.LHRP*MX9GF5>:95YIE6F6=:B8]EG@F5):95YIE6
M@F2)*95Y!I5@F6=*2>29E4IUYIE:F6=:9Y9I513_#U!+'B<4`!0````(`(5=
M=3V@#G[42B<```:2'@'?``````````C`"V@@```!()7!E<FQ];FLC9&5T
@96-T:6]N+F)M<&!+!BB`````@`!`$4``!_)P``````
`
end

.
250 2.0.0 OK 1292438925 a4si38773E8icp.155
quit
221 2.0.0 closing connection a4si38773E8icp.155


Connection to host lost.

C:\Documents and Settings\Admin>_
```

Figure 6 Advanced e-mail forging to attach files(Part B)

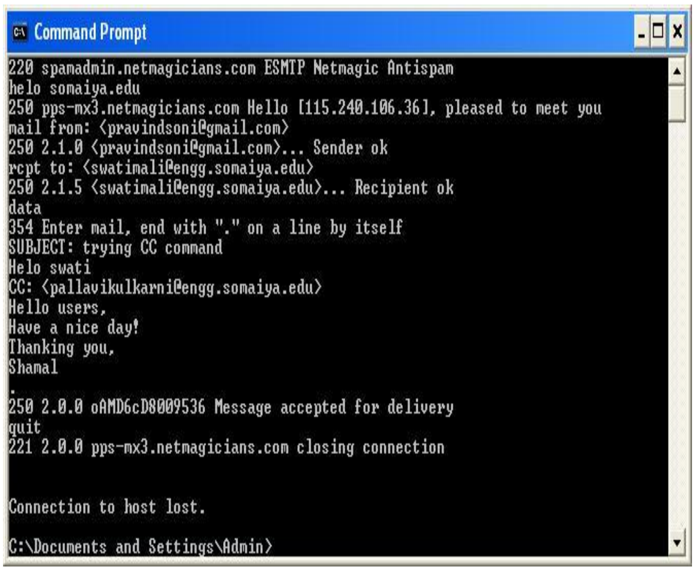
- Multiple Entries in TO Field and in CC Field

A user enters multiple e-mail addresses in both the TO field and the CC field , whenever wants to send the same e-mail to many people .The function of an entry in the CC field is equivalent to that of an entry in the TO field, even behind the scene SMTP working remains same. The e-mail addresses entered in the CC field are actually sent using multiple occurrences of the RCPT command. Example shown in fig 7 demonstrates how multiple entries can made in the CC field.



```
220 spamadmin.netmagicians.com ESMTP Netmagic Antispam
helo somaiya.edu
250 pps-mx3.netmagicians.com Hello [115.240.106.36], pleased to meet you
mail from: <pravindsoni@gmail.com>
250 2.1.0 <pravindsoni@gmail.com>... Sender ok
rcpt to: <swatimali@engg.somaiya.edu>
250 2.1.5 <swatimali@engg.somaiya.edu>... Recipient ok
rcpt to: <shamal@inbox.com>
550 5.7.1 <shamal@inbox.com>... Relaying denied. IP name lookup failed [115.240.
106.36]
rcpt to: <Manishpotey@engg.somaiya.edu>
250 2.1.5 <Manishpotey@engg.somaiya.edu>... Recipient ok
rcpt to: <deepaksharma@engg.somaiya.edu>
250 2.1.5 <deepaksharma@engg.somaiya.edu>... Recipient ok
rcpt to: <pallavikulkarni@engg.somaiya.edu>
250 2.1.5 <pallavikulkarni@engg.somaiya.edu>... Recipient ok
data
354 Enter mail, end with "." on a line by itself
subject:trying multiple recipients
helo users,
Have a nice day!
THE GOD IS GREAT
Thanking you,
shamal
.
250 2.0.0 oAMCopMX015579 Message accepted for delivery
quit
221 2.0.0 pps-mx3.netmagicians.com closing connection


Connection to host lost.

C:\Documents and Settings\Admin>_
```

Figure 7 Advanced e-mail forging to include multiple entries in TO field to send mail to multiple recipients.



```
220 spamadmin.netmagicians.com ESMTP Netmagic Antispam
helo somaiya.edu
250 pps-mx3.netmagicians.com Hello [115.240.106.36], pleased to meet you
mail from: <pravindsoni@gmail.com>
250 2.1.0 <pravindsoni@gmail.com>... Sender ok
rcpt to: <swatimali@engg.somaiya.edu>
250 2.1.5 <swatimali@engg.somaiya.edu>... Recipient ok
data
354 Enter mail, end with "." on a line by itself
SUBJECT: trying CC command
Helo swati
CC: <pallavikulkarni@engg.somaiya.edu>
Hello users,
Have a nice day!
Thanking you,
Shamal
.
250 2.0.0 oAMD6cD8009536 Message accepted for delivery
quit
221 2.0.0 pps-mx3.netmagicians.com closing connection


Connection to host lost.

C:\Documents and Settings\Admin>
```

Figure 8 Advanced e-mail forging to include CC field to send mail to multiple recipients

C. Mass Emailing

Phishers can use readymade bulk mailing tools available on net or they can build on their own. Now a days number of free softwares are there like MassMailer , BulkMail , eMailer, Mail me 3.00 etc. Most of the phishers found a lazy but efficient bulk-mailing method that does not require them to stay on the Internet while the bulk mailings are being sent. Most bulk-mailing tools are client side and require the client computer to be on the Internet while sending the e-mails. So Phishers use a PHP bulk-mailing tool that executes on the server side, which utilizes the bandwidth of the compromised dedicated server. This bulk mailing tool include four files as

1. Mail.php
2. Ini.inc
3. Maillist.txt
4. Testmail.html

Mail.php

```
<?php
include("ini.inc");
$mail_header = "From: mtechcomp2009@gmail.com";
$mail_header .= "Content-Type: text/html\n";
$subject="Account Verification Requested";
$body=loadini("testmail.html");
if (!($fp = fopen("maillist.txt", "r")))
exit("Unable to open mailing list.");
$i=0;
print "Start time is "; print date("Y:m:d H:i:s"); print "\n";
while (!feof($fp)) {
fscanf($fp, "%s\n", $name);
```

```
    print $name;
    $i++;
    mail($name, $subject, $body, $mail_header);
    }
    print "End time is "; print date("Y:m:d H:i:s");
    ?>
```

Ini.inc

The include file ini.inc, which is a header file that contains the functions we are calling within the bulkmail.php program.

```
<?php
    function loadini($path) {
    $fp = fopen($path, "r");
    $fpcontents = fread($fp, filesize($path));
    fclose($fp);
    return $fpcontents;
    }
    function readini($filename, $key) {
    return rfi($filename,$key,TRUE);
    }
    function rfi($filename, $key, $just_value) {
    $filecontents=loadini($filename);
    $key .= "=";
    $currentkey = strstr($filecontents, $key);
    if (!$currentkey)
    return($empty);
    $endpos = strpos($currentkey, "\r\n");
    if (!$endpos) $endpos = strlen($currentkey);
    if ($just_value) $currentkey = trim(substr($currentkey, strlen($key),
    $endpos-strlen($key)));
    else $currentkey = trim(substr($currentkey, 0, $endpos));
    return ($currentkey);
    }
    ?>
```

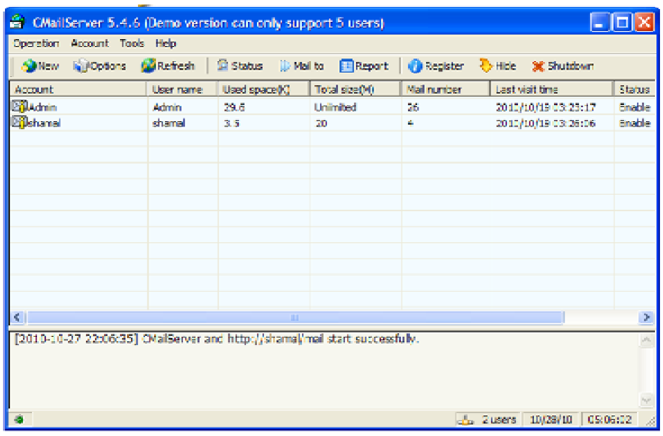
maillist.txt : The maillist.txt is a text file with the list of e-mail addresses that we plan to send to the victims.

```
admin@shamal.com
admin@shamal.com
admin@shamal.com
admin@shamal.com
```

testmail.html: The testmail.html is the e-mail we are sending.

The mail.php program has two ways of execution; via our Web browser or the command line. The command line will require us to be on the server shell and execute it, whereas with the Web browser, the phisher can hit it and exit the browser, leaving the server to do the rest of the work.

We have used CMailServer as our local mail server and run the above PHP scripts on Wamp server's localhost. Fig 8 shows that using above script we can successfully send mass emails to admin@shamal.com.


CMailServer 5.4.6 (Demo version can only support 5 users)

Account	User name	Used space(K)	Total size(M)	Mail number	Last visit time	Status
Admin	Admin	29.6	Unlimited	26	2010/10/19 03:23:17	Enable
shamal	shamal	3.5	20	4	2010/10/19 03:26:06	Enable

[2010-10-27 22:06:35] CMailServer and http://shamal/mail start successfully.

2 users 10/28/10 05:06:12

Figure 9 Mass E-mailing using CmailServer

D. Method to trace the source of an e-mail.

Each time an email is sent on the Internet, it not only carries the message body , but also transmits relevant information on the path travel by it.This information is known as Email Header of the email.Email headers are automatically generated and embedded into an email message both during composition and transfer between systems. They not only contain valuable information on the source of the email , but also represent the exact path taken by it, which can be represented as

Sender Outbox ⟶ **Source Mail Server** ⟶ **Intermediate Mail Server** ⟶ **Destination Mail Server** ⟶ **Destination Inbox**.

One can possibly identify the source of the email by simply Reverse Engineering the path traveled by it.Most Cyber Crime investigators turn to email headers for evidence in any kind of email related crime. The above email header can be divided into the following chunks:

Part A:

```
Return Path:
pankaj.chandigarh@gmail.com
Received: from
pankaj.chandigarh@gmail.com by
(208.50.6.127:25) via
   ug-out-1314.google.com
(66.249.92.171:20204)          with
[InBox.Com      SMTP      Server]id
608240002387.WM27               for
shamal@inbox.com; Thu, 24 Aug 2006
00:34:37-0800
Received: by ug-out-1314.google.com
with     SMTP     id     y2so389480uge
for <shamal@inbox.com>; Thu, 24 Aug
2006 01:34:17 -0700 (PDT)
DomainKey-Signature:     a=rsa-sha1;
q=dns;  c=nofws;s=beta;  d=gmail.com;
h=received:message-
id:date:from:to:subject:mime-
version:content-type;
b=Gq4HzLHTQwXmfvsJcJ65quwYgG9l/a6zLW
wPBr63PZ1WJE/8thjWVm+BD8GWCCg6Iu6/Cd
HYGggVpcpXqmfNO4JDtVulPMkirNemUaSltK
zjfuHF2DIJi1Zorhvq5CvxT10gTl92UjmQE5
XMZkpHqGUBSmX6O7Qwd27kfpZHjXo=
 Received:   by   10.67.119.13   with
SMTP  id  w13mr766642ugm;Thu,  24  Aug
2006 01:34:14 -0700 (PDT)
Received: by 10.66.237.1 with HTTP;
```

Part A:

Return-Path: pankaj.chandigarh@gmail.com

- This email address is the sender's email address.
- The Source Mail Server: ug-out-1314.google.com ([66.249.92.171])
- The Destination Mail Server : inbox.com
- The receiver connects to this destination mail server and download the email using simple POP command.
- Thus the complete path travelled by email can be depicted in the following manner
- (Source) 10.66.237.1-----→(Source Mail Server) ug-out-1314.google.com ([66.249.92.171]) ----→ (Destination Mail Server) inbox.com ---→TARGET SYSTEM(Destination).

Part B:

```
 Message-ID:     <20100727143839.30156.@
mail.gmail.com >Date:  Thu,  24  Aug
2006 14:04:12 +0530
 From:        "Pankaj        Mishra"
<pankaj.chandigarh@gmail.com>
 To:              shamal@inbox.com,
shamal.firake@gmail.com
 Subject:  Bjarne  Stroustrup  Book
(C++)
 MIME-Version: 1.0
 Content-Type: multipart/mixed;
 X-Spam-Ratio:           0.02------
=_Part_156068_33357618.115640845249
5
 Content-Type:
multipart/alternative;
        boundary="----
=_Part_156069_14038231.115640845249
5"

 Content-Transfer-Encoding: 7bit
```

Part B : It tells us that email was sent by pankaj.chandigarh@gmail.com

MessageId : MessageId field of the email header can be broken down in the following manner.

- 20100728080538: The email was sent in the year 2010, month July(7th),day 28th and at the time 14 hours, 38 minutes and 39 seconds.
- 30156: Each email send by the Mail Server has unique message Id reference number associated with it.The log file contain all the message Id's.

Cyber crime Investigators often use the reference number to carry out investigations.

Now up till now we have seen the basics of e-mail, details provided by e-mail header and methods used to send fake e-mails. On the basis of the above study we can give now the algorithm for e-mail tracing as follows:

1]. Algorithm For Email Tracing

- Step 1. Open the Email header.

// The SMTP protocol is used to send emails while the POP protocol is used to receive them.

- Step 2.Identify the source and destination of email by tracing the path

Sender Outbox ----→ Source Mail Server----→Intermediate Mail Server ---→Destination Mail Server ---→ Destination Inbox.

- Step 3.Identify the IP Address of the computer that was used to send the email with the help of Unique Message ID reference stored on the log file on a Mail Server. **OR**

Step 3 can be performed as below

// Use Reverse DNS look up ie convert the suspected IP Address into the corresponding hostname.

Use utility named *nslookup.*

$>nslookup IP Address of the sender

$>nslookup 203.94.243.71

203.94.243.71 has valid reverse DNS of mail2.mtnl.net.in

OR

Step 3 : Write a program to convert the IP Address to hostname or vice versa using JAVA coding IPAddress API

2]. Method to Trace forged E-mail

The above algorithm gives us the physical source of an e-mail , provided it is sent by an authorized user. To trace a fake e-mail sent by using e-mail forging or mass e-mailing we need more details again. To trace these e-mails following steps can be followed.

- Check the final received header field , because it can not be forged in any case. So if the DNS names are different in sender's e-mail address and final received header that means it is a forged e-mail.
- Now to trace it one can examine tcp_wrapper, ident, and sendmail logs to obtain information on the origin of the spoofed email.
- The header of the email message often contains a complete history of the "hops" the message has taken to reach its destination. Information in the headers (such as the "Received:" and "Message-ID" information), in conjunction with your mail delivery logs, should help you to determine how the email reached your system.
- If your mail reader does not allow you to review these headers, check the ASCII file that contains the original message.

This process may help you to trace a forged or fake e-mail but realize that in some cases, you may not be able to identify the origin of the spoofed email.

IV. Proposed Model To Detect And Prevent Phishing Attacks

Now a days phishing e-mail attacks are very easy for fradulents to carry out. As mass e-mails are sent , the number of affected victims are also large. To fight against such attacks, we proposed an anti-phishing tool to detect and prevent e-mail phishing attacks.

Problem Statement :

Detection and Prevention of Phishing Attacks on Email.

Fig. 9 Shows the basic architecture of the tool.

A. Modules of the Application

The tool mainly implements following modules

1. User Interface Module

User friendly graphical interface will be developed by using java technology for ease of use. It facilitates the use of tool for naive users also.

.

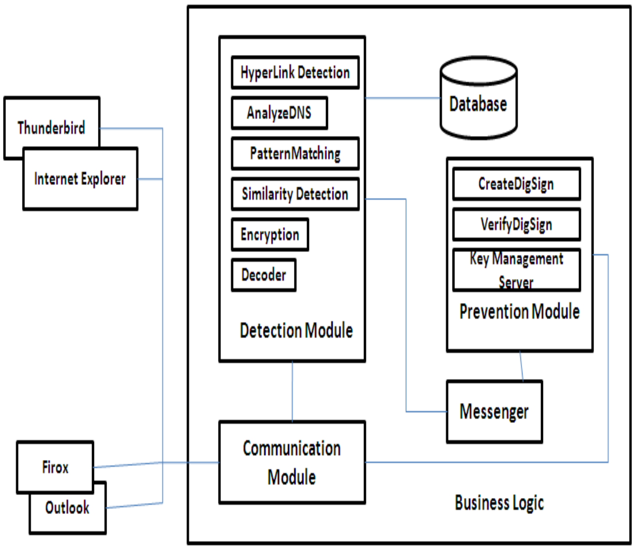


Figure 10 Arhitecture of the tool to detect and prevent phishing attacks

2. Database Maintenance Module

Data storage module storing, managing and if needed update the URL and IP address information of trusted websites

3. Business Logic Module

This module implements Hyperlink Detection Module. It uses data provided by user emails and database. It contains sub modules as

- Detection Module
- Prevention Module
- Communication Module
- Messenger Module

3.1 Detection Module

Detection Module reads the mails from inbox of mail client of user. It scans all messages and detects for any phishing attack , by using generic characteristics of

Hyperlinks. The Detection Module includes following sub-modules,

a. Hyperlink Detection Module

Hyperlink Detection Module fetches the DNS names of actual link and visual links of hyperlinks. If both the links are not empty and are different then it warns user about the phishing attack. Again it checks whether the actual DNS name is directly used as dotted decimal then returns possible phishing attack. Many times to confuse the user the actual links and visual links are encoded by using Hexadecimal code or ASCII code. To handle such situations module calls the respective DECODER modules and then compare the decoded links. Module also checks the JavaScript attack if any present in an email. It just checks for the keyword "Java Script" in the email text and if it present, module warns user as possible phishing Attack. This module implements DetectHyperLink Algorithm.

b. AnalyzeDNS Names Module

AnalyzDNS Names module is used if visual link is null in the hyperlink. The module then check the DNS of hyperlink in Blacklist and whitelist respectively. If It doesn't find there also, then it calls the pattern matching module.This module is implemented using AnalyzeDNS algorithm.

c. PatternMatching Module

It implements the PatternMatching algorithm. Pattern matching module first extracts the DNS name from sender's email address.If this senders DNS name and actual link DNS name are different then it is possibly a phishing attack.If both are same then module checks the previously accessed links database maintained as SEED_SET .SEED_SET is a list of possible phishing links previously accessed or identified. Module then checks the DNS name of actual link against each item in the SEED_SET. If match is found module returns as POSSIBLE PHISHING attack .To compare it calls the Similarity algorithm module.

d. Similarity Detection Module

This module checks how much similar is the actual link DNS name with an item in the SEED_SET. If similarity is beyond a threshold value then it returns true otherwise false. This module uses Similarity algorithm.

e. Encryption Module

It Implements MD5 algorithm to calculate message digest of URL and IP addresses of such institutions/websites where he sends his login details, i.e., username and password. **MD5** (**Message-Digest algorithm 5**) is a widely used cryptographic hash function with a 128-bit hash value. MD5 is commonly used to check the integrity of files and has been employed in a wide variety of security applications.

f. DECODER Module

This module consist of two parts as

- **Hexadecimal Decoder**

Many times the hyperlinks are mentioned in Hexadecimal format so that normal user may get confused. To understand the actual DNS name of encoded hyperlink , it must be first decoded. Hexadecimal Decoder algorithm decodes the given Hexadecimal given format into normal text.

- **ASCI I Decoder**

Likewise , the hyperlinks are mentioned in Hexadecimal format also so that normal user may get confused. To understand the actual DNS name of encoded hyperlink , it must be first decoded. The algorithm decodes the given ASCII given format into normal text.

3.2 Prevention Module

Prevention Module helps user to prevent from phishing attacks. It allows user to create Digital Signatures to send the official messages .The receiver will verify the Digital Signature at other end and authenticates the sender of the message. A message signature is essentially a sophisticated one-way hash value that uses aspects of the sender's private key, message length, date and time. In general the module does the following things

- Create a personal public/private key pair

Upload their public key to respected key management servers so that other people who may receive emails from the user can verify the messages integrity.

- Enable, the automatic signing of emails

Verify all signatures on received emails and be careful of unsigned or invalid signed messages – ideally verifying the true source of the email

3.3 Communication Module

Communicate with all of the monitored processes, collect data related to user input from other processes (e.g. IE, outlook, firefox, etc.), and send these data to the Business Logic module, it can also send commands (such as block the phishing sites) from the Business Logic executive to other modules.

3.4 Messenger

When receiving a warning messages from Business Logic module , it shows the related information to alert the users and send back the reactions of the user back to the Business Logic module.

V. CONCLUSION

The specter of online identity threat was never so real as it is today primarily due to rapid growth of the Internet and increase in online trading activities which offer a cost effective method to service providers, such as banks, retailers etc., to reach out to their customers via this medium. This has also provided the phishing community an excellent tool to try and fool the netizans into divulging sensitive information about their banking accounts, credit cards details, etc. Recent years have witnessed a host of phishing scams with each doing the other in terms of reach to the users and the level of sophistication.

Though the best measure available against such scams is user awareness , it is highly impossible also. So many tools have been developed to fight against the e-mail phishing attacks. To contribute in this regard we, have also taken a step ahead. This paper gives the details literature survey of the approaches till now used by different people to detect and prevent e-mail phishing attacks. We have given the details of how e-mail phishing attacks are carried out with experimentation results. We have also proposed our own approach to fight against the e-mail phishing attacks. The modules include the detail specification of their functionality. Thus , we assure that this solution will help to the normal end user as well as the corporate people also to send the highly confidential data.

REFERENCES

1. The Anti-phishing working group. http://www.antiphishing.org/.
2. A. Williams. “Phishing Exposed”. Syngress Publishing Inc.; 2005.
3. I. Fette, N. Sadeh, and A. Tomasic. “Learning to detect phishing emails”. Technical Report CMU-ISRI-06-112, Institute for Software Research, Carnegie Mellon University, June 2006. http://reports-archive.adm. cs.cmu.edu/anon/isri2006/abstracts/06-112.html
4. Juan Chen and Chuanziong Guo “Online Detection and Prevention of Phishing Attacks” IEEE Communications and Networking, ChinaCom ’06, pp 1-7, Oct 2006.
5. Ollmann, G., “The Phishing Guide”. NGS Software Insight Security Research 2005, http://www.ngs software.com/papers/NISRWPhishing.pdf.
6. Engin Kirda and Christopher Kruegel, “Protecting Users Against Phishing Attacks” in Computer Software and Applications Conference, 2005 (COMPSAC 2005), Edinburgh, Scotland. 29th Annual International Volume 1, pp. 517 – 524, Issue: 26-28 July 2005.
7. M. Chandrashekaran, K. Narayana, S. Upadhyaya,“Phishing Email Detection Based on Structural Properties”, s*ymposium on Information Assurance: Intrusion Detection and Prevention*,New York, 2006
8. R. Suriya1 , K. Saravanan2 and Arunkumar Thangavelu An Integrated Approach to Detect Phishing Mail Attacks A Case Study, 3, *SIN’09,* October 6–10, 2009, North Cyprus, Turkey. Copyright 2009 ACM 978-1-60558-412-6/09/10
9. Phishing in Finance, Kristofer Beck, Justin Zhan, 978-1-4244-6949-9/10/$26.00 ©2010 IEEE
10. Huajun Huang, Shaohong Zhong, Junshan Tan, Browser-side Countermeasures for Deceptive Phishing Attack, 2009 Fifth International Conference on Information Assurance and Security
11. Danesh Irani, Steve Webb, Jonathon Giffin and Calton Pu,” Evolutionary Study of Phishing” 978-1-4244-2969-1/08/ c_ 2008 IEEE
12. Jordan Crain, Lukasz Opyrchal, Atul Prakash,”Fighting Phishing with Trusted Email”**,** 2010 International Conference on Availability, Reliability and Security
13. B. Adida, S. Hohenberger, and R. L. Rivest, “Fighting phishing attacks: a lightweight trust architecture for detecting spoofed emails,” February 2005, draft.
14. P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In CHI ’07: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 905–914, New York, NY, USA, 2007. ACM.
15. Wilfried N. Gansterer, David P¨olz,” E-Mail Classification for Phishing Defense”.
16. Kapil Oberoi and Anil K. Sarje , “An Anti-Phishing Application for the End User” 3rd Hackers’ Workshop on Computer and Internet Security March 17-19, 2009, Prabhu Goel Research Centre for Computer & Internet Security Department of Computer Science and Engineering Indian Institute of Technology Kanpur
17. Weider D. Yu Shruti Nargundkar Nagapriya Tiruthani,” **PhishCatch – A Phishing Detection Tool”,** 2009 33rd Annual IEEE International Computer Software and Applications Conference.
18. Phishing in Finance, Kristofer Beck, Justin Zhan, 978-1-4244-6949-9/10/©2010 IEEE
19. Phishing Activity Trends Report, 2009, Available online: http://www.antiphishing.org/reports/apwg_report_ 2009.pdf
20. Phishing Activity Trends Report, 1st Half 2010, Available online: http://www.antiphishing.org/reports/apwg_report_h1_2010.pdf
21. Mohamad Badra, Samer El-Sawda, Ibrahim Hajjeh,” Phishing Attacks and Solutions**”** *Mobimedia'07*, Month 8, 2007, Nafpaktos, Aitolokarnania, Greece. Copyright 2007 ICST 978-963-06-2670-5
22. *Soroush Dalili,”* How to prevent phishing attacks?- In 3 Pages -
23. Sun Bin, Wen Qiaoyan, Liang Xiaoying,” A DNS based Anti-Phishing Approach”, 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing
24. David Harley,”A Preety Kettle Of Phish”.
25. “Today’s Blended Threats”, White Paper, Symantec Internet Security Threat Report, Trends for July-December 06, p. 13.
26. Tod Beardsley,”Phishng Detection and Prevention ,A practical Counter Fraud Solutions”.
27. “Discover What Your Boss Is Looking At”
28. Lightweight Signatures for Email By Ben Adida_ David, Chau_ Susan ,Hohenberger_,† Ronald L. Rivest, Computer Science and Artifical Intelligence Laboratory.
29. Phishing Secrets: History, E®ects, and Countermeasures, Antonio San Martino and Xavier Perramon, *International Journal of Network Security, Vol.11, No.3, PP.163{171, Nov. 2010*

30. Putting an End to Account-Hijacking Identity Theft Federal Deposit Insurance Corporation ,Division of Supervision and Consumer Protection ,Technology Supervision Branch ,December 14, 2004

31. Phishing attacks and Countermeasures, Anil Sagar,Operations Manager

32. Anti-Phishing *Best Practices for Institutions and Consumer,Mccafe*

33. A Practical Approach to Managing Phishing Michael Barrett, Chief Information Security Officer Dan Levy, Senior Director of Risk Management – EuropeApril 2008

34. “E-MAIL HACKING” By Ankit Fadia.

A Study on Searching Mechanisms in Semantic Web

M.Thangaraj[1] and G.Sujatha[2]

[1]Madurai Kamaraj University, Department of Computer Science,*thangarajmku@yahoo.com*
[2]Sri Meenakhsi Govt. College For Women (A), Department of Computer Science.

[2]Corresponding Author sujisekar05@rediffmail.com

Abstract: The internet presents a huge amount of useful information which is usually formatted for its users which makes it difficult to retrieve relevant data from various sources. With the growing complexity of online information the search results in keyword based search engine are growing increasingly vague and cumbersome. The existing information retrieval systems are mostly keyword-based and retrieve relevant documents or information by matching keywords. Keyword-based search in spite of its merits of expedient query for information and ease-of-use has failed to represent the complete semantics contained in the context and has led to the retrieval failure. Although many approaches for Information Retrieval in semantic web has been developed, there has been limited effort to compare such tools. The architectural aspects of a few semantic search systems were presented by comparing various features.

Keywords: Semantic Web, Semantic search, Information Retrieval, Ontology, Search Engine, Semantic Similarity.

1. Introduction

The semantic web [6] is an extension of the current Web in which resources are described using logic-based knowledge representation languages for automated machine processing across heterogeneous systems. In recent years, its related technologies have been adopted to develop semantic-enhanced search systems.

Semantic Search Systems (SSS) are Information Retrieval (IR) Systems that employ semantic technologies to enhance different parts of IR by using semantic Relations, Ontologies, Clusters, Crawlers and Similarities. Research in IR community has developed variety of techniques to help people locate relevant information in large document repositories.

Besides classical IR models i.e., Vector Space and Probabilistic Models[4] extended models such as Latent Semantic Indexing, Machine Learning based models i.e., Neural Network, Symbolic Learning, and Genetic Algorithm based models and Probabilistic Latent Semantic Analysis (PLSA) have been devised with hope to improve information retrieval process. However, rapid expansion of the Web and growing wealth of information pose increasing difficulties to retrieve information efficiently on the Web. To arrange more relevant results on top of the retrieved sets, most of contemporary Web search engines utilise various ranking algorithms such as PageRank, HITS, and Citation Indexing that exploit link structures to rank the search results. Despite the substantial success, those search engines face perplexity in certain situations due to the information overload problem on one hand, and superficial understanding of user queries and documents on the other.

Significance of the research in this area is for two reasons: it supplements conventional information retrieval by providing search services centered on entities, relations, and knowledge; and development of the semantic web also demands enhanced search paradigms in order to facilitate acquisition, processing, storage, and retrieval of the semantic information. This paper provides a survey to gain an overall view of the current research status in this area. We classify our studied systems into several categories according to their most distinctive features, as discussed in the next section. The categorization by no means prevents a system from being classified into other categories. We provide a review focusing on objectives, methodologies, and most distinctive features of individual systems; and discuss issues related to knowledge acquisition and search methodologies.

In this paper, We focus on Semantic Search architectures from five directions. They are i. Relation Centered Search ii . Ontology Centered Search iii. Similarity Based Search iv.Crawler Based Search v. Cluster Based Search. This paper is organized as follows. Section 2 introduces related work in this area. Then the semantic search directions are presented in Section 3. Finally the conclusions are made in Section 4.

2. Semantic Search Systems

The unsolved problems of current search engines have led to the development of the semantic web search systems [31]. Search is one of the most popular applications on the web and it is an application with significant room for improvement. The addition of explicit semantics can improve search. Semantic search attempts to augment and improve traditional search results by using data from the semantic web [10].

Variety of SSS consists of different tools: semantic browsing with automatically generated annotations, Semantic Query expansion, Semantic Ranking, Systems working on a Single Ontology or Multiple Ontologies. There exist various attempts to classify the searching system. For instance, distinguish four key characteristics of semantic metadata based search systems: search environment, query type, intrinsic problems, iterative and exploratory dimensions.

Furthermore, the SSS are classified by semantic technology usage, and the usage of ontology and its elements. We summarize important categories [7] in Fig. 1 based on analysis of the literature and related classification schemes.

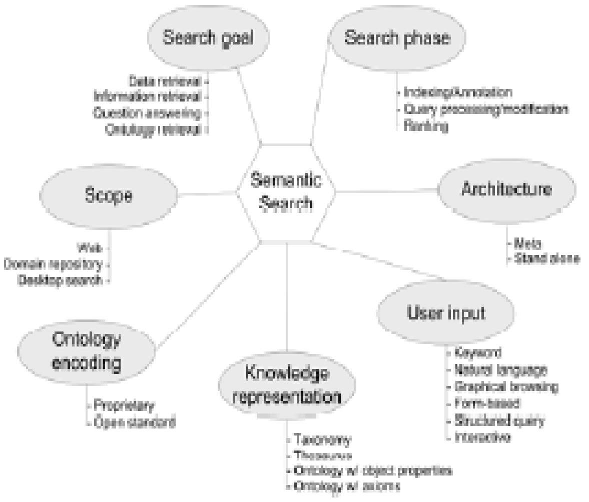


Fig 1: Classification of semantic search systems

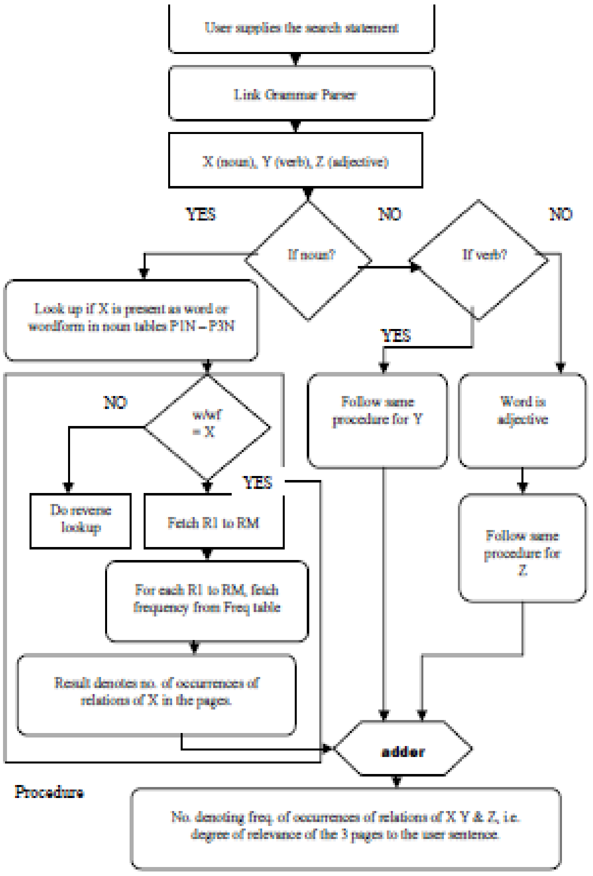

User supplies the search statement
Link Grammar Parser
X (noun), Y (verb), Z (adjective)
YES
NO
NO
If noun?
If verb?
Look up if X is present as word or wordform in noun tables P1N – P3N
YES
NO
w/wf = X
YES
Do reverse lookup
Fetch R1 to RM
For each R1 to RM, fetch frequency from Freq table
Result denotes no. of occurrences of relations of X in the pages.
Follow same procedure for Y
Word is adjective
Follow same procedure for Z
adder
Procedure
No. denoting freq. of occurrences of relations of X Y & Z, i.e. degree of relevance of the 3 pages to the user sentence.

Fig 2. Algorithm to Perform Related Keyword Search

3. Directions in Semantic Search System

The Semantic Search architectures from five directions i.Relation Centered Search ii. Ontology Centered Search iii.Similarity Based Search iv.Crawler Based Search v. Cluster Based Search are discussed in this section.

3.1 Relation-Centered Search

Relation-Based search is an extension of the conventional IR approaches where the main goal is to retrieve the most meaningful pages only. In this type of search system the retrieval process is carried out by matching user queries with the relationship between keywords.

The first approach for ranking the pages is based on the content count for a particular keyword [3]. For the given noun N, the frequency of occurrence of N in the database is keyword searching, finding the frequency of occurrence for its relations in that page gives us a count of how meaningful the page is towards N. The page which has a highest number of related keyword hits a higher page rank for the user entered keyword. Consider the three web documents, the parts of speech, such as nouns, verbs and adjectives were extracted using a grammatical parser like Link Grammar Parser. These parts were then fed to a lexical dictionary WordNet, to extract the various relations such as synonyms, hypernyms, hyponyms, meronyms, and holonyms. These relations were then stored in tables categorized by their part of speech, for e.g. P1N, P2N, P3N, P1ADJ, P2ADJ, P3ADJ, P1V, P2V, and P3V, where P1N stands for "nouns on page1 with its extracted relations", P1V stands for "verbs on page1 with its extracted relations", and so on. Now the user query was processed by an algorithm depicted in Fig. 2.

The Nouns, Verbs, Adjectives (X, Y, Z) were first extracted from the user queried sentence. The noun X, was first checked in the three tables of nouns for the three web pages. If X was present physically in any of the three web pages, would get a hit in the "word" column or "word form" column of the three noun tables P1N-P3N. If there is a hit, consider the corresponding relations for noun X. Then perform a frequency search for those relations of X in the three pages. So in this way, the system searched for related keywords or keyword relations in a webpage. This count gave an idea of how relevant the webpage was towards the keyword. Similar treatment was allotted to the other keywords of the user queried sentence. The total count got was a measure of the page relevance towards the user queried sentence. This work is based on the relation count between the keywords. It was improved by adding semantic meanings in the SemSearch System.

"SemSearch" hides the complexity of semantic search from end users and to make it easy to use and effective for novice users [33]. SemSearch is a layered architecture (Fig.3) that separates end users from the back-end heterogeneous semantic data repositories. User Interface Layer, allows end users to specify queries in terms of keywords. The Text Search Layer makes sense of user queries by finding out the explicit semantic meanings of the user keywords. Two components namely a semantic entity index engine (indexes documents and their associated semantic entities including classes and properties) and semantic entity search engine (supports the searching of semantic entity matches for the user keywords) are central to this layer.

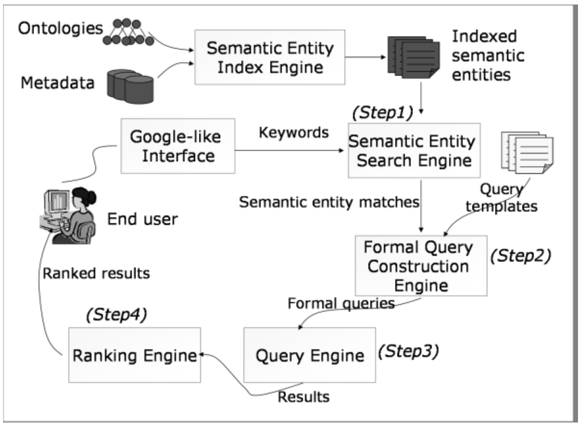


Fig. 3. An overall diagram of the SemSearch search engine

The semantic query layer produces search results for user queries by translating user queries into formal queries which comprises of a formal query construction engine (provides a specific formal query language that can be used to retrieve semantic relations from the underlying semantic data layer), a query engine (queries the specified metadata repository using the generated formal queries) and a ranking engine (ranks the search results according to the degree of satisfaction on the user query). The semantic data layer, comprises of semantic metadata that are gathered from heterogeneous data sources and are represented in different ontologies. SemSearch accepts keywords as input and produces results which are closely related to the user keywords in terms of semantic relations. This method is modified by formulating concept based keywords.

The early work [34] “Ontolook”, is a relation based search engine provides the relationship between the keywords in terms of the concepts. Initially “OntoLook” will analyze the keyword combination input by the user. The system will analyze these inputs and handle the inherited relationship between these concepts. Then, these concepts are assembled to some concept pairs and send these pairs to the ontology database to retrieve all relations defined by ontology between concept pairs.

After all relations between concept pairs are retrieved from the ontology database, a concept-relation graph is formed based on these relations and concepts. Then “OntoLook” will cut less relevant arcs from the graph. If the number near the arc is larger then it denotes the maximum relations between the concepts. Otherwise if the number near the arc is zero, the algorithm behaves like a Keyword based search. Because there are some relations between the keywords which user input, the result set retrieved from the database will be close to the users’ intention when less-ranked arcs were cut from the graph. Finally, the system fetches the relation and its corresponding keyword pair from each arc in sub graphs to form a property-keyword candidate set. Then, the property keyword candidate set is sent to the database to get a retrieved result set for the users.

In this architecture (Fig 4) a crawler program collects the web pages on the internet with its semantic markup and corresponding ontology which is described in an OWL document in the Internet. The collected web pages are transported to a web page database to be stored for the use of future retrieving URLs and corresponding web pages. The ontology, OWL document, is conveyed to an OWL Parser. The OWL parser will map the ontology into a relational database.

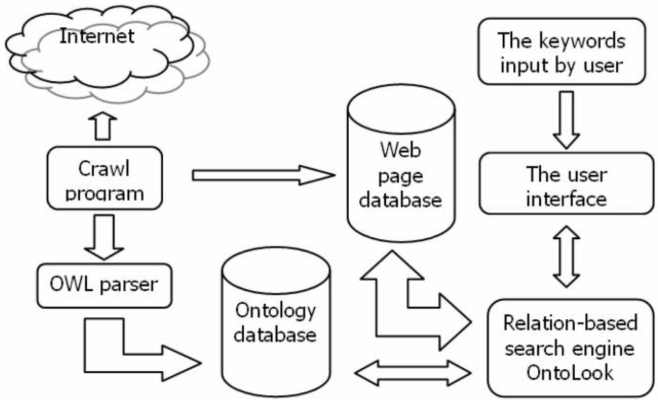


Fig. 4. System architecture of “OntoLook.

The effectiveness of this approach is limited by lack of priority ranking technology and page rank technology to make a relation based page rank. With this considerations the next work is presented in [9] by providing effective page ranking. The Annotated Web pages from the SemanticWeb including RDF metadata are collected by the crawler application and originating OWL ontology. The OWL Parser interprets the RDF metadata and stored in the knowledge database. A graphics user interface allows for the definition of a query, which is passed on to the relation-based search logic. The ordered result set generated by this latter module is finally presented to the user.

The “ranking criterion” (Fig 5) is based on the estimate of the probability that keywords/concepts within an annotated page are linked one to the other in a way that is the same to the one in the user’s mind at the time of query definition. This probability measure can be effectively computed by defining a graph-based description of the ontology (ontology graph), of the user query (query subgraph), and of each annotated page containing queried concepts/keywords (both in terms of annotation graph and page subgraph).

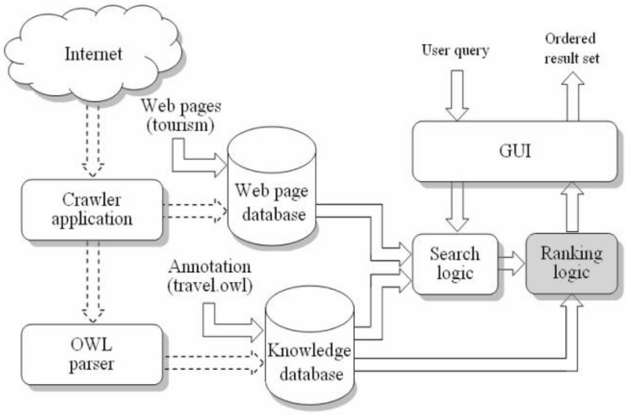


Fig 5. Semantic Web infrastructure (prototype architecture).

Given an ontology graph and a query subgraph a ranking strategy is designed. This strategy will assign a relevance score to each page including queried concepts based on the semantic relations. As per the proposed ranking strategy, for the given query Q, for each page p, a page subgraph can be built and exploiting the information available in page annotation. The methodology starts from a page subgraph computed over an annotated page and generates all the possible combinations of the edges belonging to the subgraph by excluding cycles.

There may be pages in which there are concepts that do not show any relations with other concepts. But that could still be of interest to the user. The methodology progressively reduces the number of edges in the page subgraph. Then it computes the probability of the resulting subgraphs obtained by a combination of the remaining edges that matches the user's intention. Edge removal could lead to having concepts without any relation with other concepts. Thus, several relevance classes are defined, each characterized by a certain number of connected concepts. Within each class, pages are ordered depending on the probability measure above and presented to the user. An enhancement of present work with multiple ontologies is seemed to be effective.

3.2 Ontology-Centered Search

Ontology provides a flexible way of introducing semantics into the semantic web. The main advantage of using ontologies is reusability of knowledge. A number of ontology libraries currently exist. Example libraries are Ontolingua (www.ksl.stanfor.edu/software/ontolingua) and OWL library (http://protege.stanford.edu/plugins/owl/owl-library). To get the right information, the search engines must be capable of finding the suitable ontologies. Some ontology based search engines available currently are Swoogle, Ontosearch.

In Ontosearch [32] which combines the google search engine together with the RDFs ontology (hierarchy) visualization technology. It will search for relevant (based on keywords) ontology files on the Internet and displays the files in a visually appealing way—as a hierarchy tree. The hierarchical view allows the users to quickly review the structures of different ontology files and select the suitable ontology files.(Fig.6)

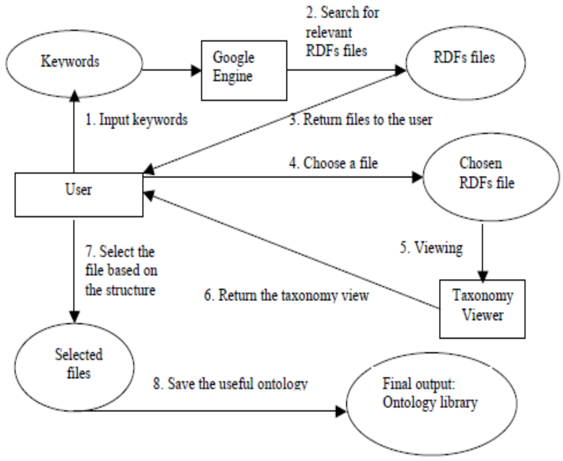

Keywords
Google Engine
2. Search for relevant RDFs files
RDFs files
1. Input keywords
3. Return files to the user
User
4. Choose a file
Chosen RDFs file
5. Viewing
7. Select the file based on the structure
6. Return the taxonomy view
Taxonomy Viewer
Selected files
8. Save the useful ontology
Final output: Ontology library

Fig 6. Overview of OntoSearch

The user inputs the keywords to describe the nature of the required ontology to OntoSearch. Then OntoSearch applies the Google engine to search for RDFs files related to the keywords and returns a list of relevant links (URLs) to the user. The user then chooses some of the returned RDFs files and displays their structure, and decides which of the files are relevant. Finally, the user select the relevant RDF files and saves them in a taxonomy library.

Ontology-searching tool OntoSearch, can be linked to the tool Information Knowledge Base (IKB). Fig 7 discusses links between them and demonstrates how they interoperate for future use.

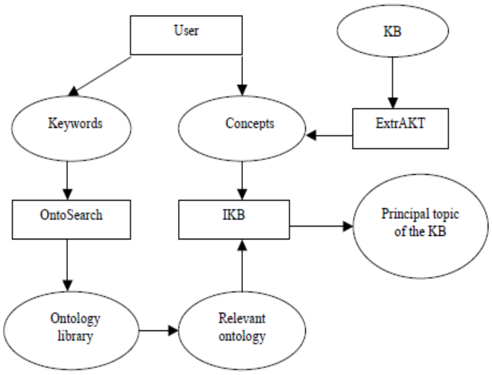


Fig 7. The relation between OntoSearch, IKB and ExtrAKT

The improved version of the previous work is based on Ontology-Based Knowledge Base by using vector space.

In the literature of semantic search engine based on Ontologies [28], the traditional Term Document Matrix (TDM) is extended to reflect the relevance between Ontologies, Web documents and terms. This extension of the traditional Vector Space Method (VSM) with semantic support. The search process begins with the parsing of a user's query (Fig 8). If a search request is in form of keyword list, then these keywords would be first treated as concepts in ontology, and documents that relates to these concepts will be retrieved based on the extended TDM. Through a user interface, a user can also submit requests by using a search wizard where user is given advanced options for a query. These options may include the ontology server, premises, answer patterns, maximum number of answers, and so on. In either forms, the request will be parsed into OWL-QL and then sent to the Reasoner which will return a set of RDF (Resource Description Framework) [36] triples containing qualified concepts/individuals in domain ontologies. After that, a document retriever finds all documents that are relevant to these concepts/individuals, and these documents are sorted by a ranker based on the relevance to the search request before they are presented to the user.

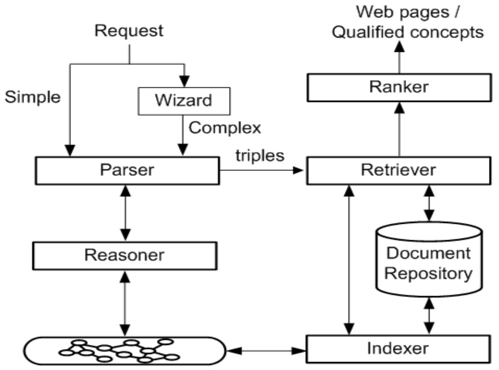


Fig 8. Query processing

The modification of the previous work is in [17] as the exploitation of ontology-based knowledge bases to improve search over large document repositories. This approach deals with an ontology-based scheme for the semiautomatic annotation of documents and a retrieval system. The retrieval model is based on an adaptation of the classic vector-space model, which includes an annotation weighting algorithm, and a ranking algorithm.

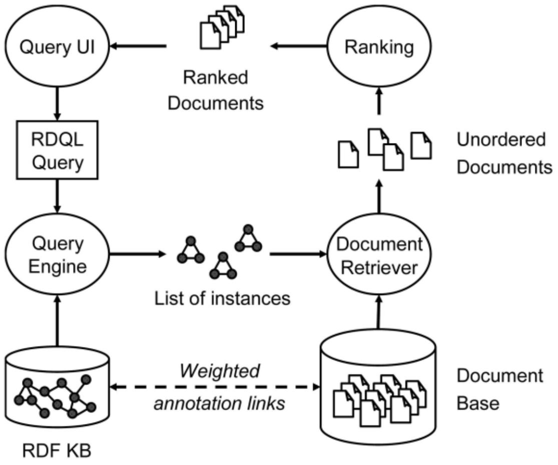


Fig 9. Overview of Ontology Based IR

This approach can be viewed as an evolution of classic keyword-based retrieval techniques, where the keyword-based index is replaced by a semantic knowledge base. The overall retrieval process is illustrated in Fig.9 that consists of the following steps: The input to the system is a formal RDQL query. The RDQL query is executed against the knowledge base, which returns a list of instance tuples that satisfy the query. This step of the process is purely Boolean (i.e., based on an exact match), so that the returned instances must strictly hold all the conditions in the formal query. Finally, the documents that are annotated with the instances returned in the previous step are retrieved, ranked, and presented to the user. The efficiency of this method is improved by using inverted list indexing structure in Ontology Knowledge Bases.

The Ontology based Information Retrieval System uses inverted tables [35]. A new third layer in the existing 2-layer inverted list is introduced for storing the ontology terms belonging to the corresponding keywords. The architecture of the system consists of two parts: the information storage part (runs background and offline) and the query part (query runs instant and online). The ontology terms of query is corresponding to the terms in the inverted files, which could improve precision of the system.

The previous work is modified with the help of semantic annotations [12]. A semantic expansion search is proposed based on constructed domain ontology, semantic annotation algorithm and semantic expansion reasoning algorithm. The experimental results show that this methodology can overcome limitations in comparison with traditional keyword search mode, and achieve higher recall ratio and precision ratio.

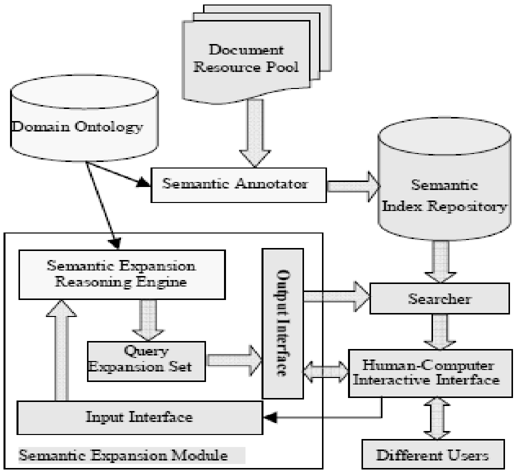


Fig.10. Semantic Expansion Search Model

Semantic expansion search model Sem-Exp-M is shown in Fig. 10. The function of semantic expansion module is to implement semantic expansion for user's query keyword. By the acquisition of search condition from human-computer interactive interface, reasoning engine executes reasoning and generates query expansion set via semantic expansion reasoning algorithm. Semantic annotator is to convert document resource pool with semantic feature. Searcher acquires query expansion set as search condition from output interface and retrieves documents from semantic index repository. This work is improved by introducing logic reasoned in the next model.

The logical reasoning based information retrieval model for the Semantic Web [24], uses OWL Lite as standard ontology language. The terms defined in ontology are used as metadata to markup the Web's content; these semantic markups are semantic index terms for information retrieval. The equivalent classes of semantic index terms by using description logic reasoner can be obtained. The logical views of documents and user information needs, generated in terms of the equivalent classes of semantic index terms. The performance of information retrieval can be improved effectively when suitable ranking function is chosen. Fig 11.

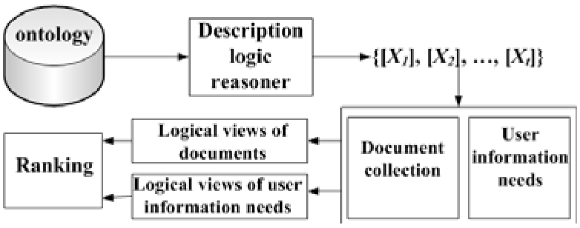

ontology
Description logic reasoner
{[X_1], [X_2], ..., [X_l]}
Logical views of documents
Ranking
Logical views of user information needs
Document collection
User information needs

Fig 11. Key parts of ontology-based information retrieval

The next work presents a methodology for the ontology based semantic annotation of web pages with annotation weighting scheme [25]. The retrieval model is based on the importance factors of the structural elements, which are used to re-rank the documents retrieval by the ontology based distance measure. The relevance concept similarities are combined with the annotation-weighting scheme to improve the relevance measures. A number of annotation tools for producing Semantic markups exist such as SHOE, Protégé, OntoAnnotate and MnM [11] [13].

The previous work is improved by adding ranking for ontologies. The study of "Ranking Ontologies Based on OWL Language Constructs" [20] uses more than one ontology to get the right kind of information the user is looking for. To present the suitable ontology to the user the ontologies are ranked by measuring two scores such as 1) How well the concept is described in terms of OWL constructs in a particular relevant class? 2) How much portion of the given ontology has the relevant OWL classes that describe the concept the user is looking for? Ontoweight is calculated by the Ontology Ranking Engine. The ontology that has the highest Ontoweight score will be ranked first.

3.3 Similarity Based Search

Similarity ranking is a hot topic in database research. Determining the semantic similarity is an important issue in the development of semantic search technology. An approach to determine the semantic similarity [15] between two entities that reflects in context. The semantic ranking approach assigns a value to the total number of entities and relations that match a user's interests.

The ranking score is defined as a function of some particular parameters. An Approach to Determine Semantic Similarity (ADSS) combines the Tabu Search algorithm with an efficient multiobjective programming algorithm to improve precision. Aleman-Meza et al [2] discuss a framework that uses ranking techniques to identify more interesting and more relevant semantic associations and define a ranking formula that considers subsumption weight, path length weight, and context weight and trust weight for assessing the effectiveness of the ranking scheme outlined. Rodriguez and Egenhofer [22] present an approach to computing semantic similarity across different ontologies.

A similarity function determines similar entity classes by using a matching process over synonym sets, semantic neighborhoods, and distinguishing features. In the SWAP project, Broekstra et al. [5] aim at overcoming the lack of semantics by combining the Peer-to-Peer paradigm with Semantic Web technologies. They propose a data model for encoding semantic information that combines ontology features with a flexible description and rating model. In Rodriguez and Egenhofer's approach, three ideas are presented—word matching, feature matching, and semantic-neighborhood matching. Broekstra et al. extend Rodriguez and Egenhofer's approach with a fourth idea—instance matching.

Thus, two objects can be identified through these similarity measures. Pekar and Staab [18] address the problem of automatically enriching a thesaurus by classifying new words into its classes. The proposed classification method uses the distributed data about a new word and the strength of the semantic relatedness of its target class to the other likely candidate classes. In contrast to the above work, ADSS introduces a multiobjective programming algorithm to compute the weights and the Tabu Search to compute the optimal solution. Hence the approach can acquire the results with higher precision. This method is modified with the heuristic mapping method in the next work.

One of the literature named "Ontology Mapping for information retrieval" [30] deals with a heuristic mapping method and a prototype mapping system that support semi-automatic ontology mapping for improving semantic interoperability in heterogeneous systems. This approach (Fig 12) is based on the idea of semantic enrichment, i.e., using instance information of the ontology to enrich the original ontology and calculate similarities between concepts in two ontologies. This approach consists of two phases: enrichment phase and mapping phase. The enrichment phase is based on analysis of the extension information in the ontologies.

The extension made in this work is written documents that are associated with the concepts in the ontologies. The intuition is that given two to-be-compared ontologies, construct representative feature vectors for each concept in the two ontologies. The documents are ''building material'' for the construction process, as they reflect the common understanding of the domain. Outputs of the enrichment phase are ontologies with feature vector as enrichment structure. The mapping phase takes the enriched ontology and computes similarity pair wise for the element in the two ontologies. The calculation is based on the distance of the feature vectors. Further refinements are employed to re-rank the result via the use of WordNet.

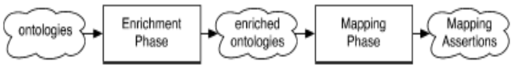


Fig. 12. Two phases of the whole mapping process

The previous work is improved in the next work by introducing Kolmogorov complexity. "The Google Similarity Distance" [23] deals with words and phrases acquire meaning from the way they are used in society, from their relative semantics to other words and phrases. A new theory of similarity between words and phrases based on information distance and Kolmogorov complexity is presented. The World Wide Web (WWW) is treated as the database, and Google as the search engine. The method is also applicable to other search engines and databases. This theory is then applied to construct a method to automatically extract similarity, the Google similarity distance, of words and phrases from the WWW using Google page counts. This model is improved by introducing a similarity ranking in literature[26].

"Scalable Probabilistic Similarity Ranking" is a scalable approach for probabilistic top-k similarity ranking on uncertain vector data. Each uncertain object is represented by a set of vector instances that is assumed to be mutually exclusive. The objective is to rank the uncertain data according to their distance to a reference object. The proposed framework computes instance and ranking position for each object, the probability of the object falling at that ranking position. The resulting rank probability distribution can serve as input for several state-of-the-art probabilistic ranking models. Existing approaches compute this probability distribution by applying the Poisson binomial recurrence technique of quadratic complexity. This complexity is reduced to a linear-time complexity with the same memory requirements in this framework. It is facilitated by incremental accessing of the uncertain vector instances in increasing order of their distance to the reference object.

3.4 Cluster Based Search

Search results on the Web are traditionally presented as a flat ranked list of documents. The main use for clustering is not to improve the actual ranking, but to give the user a quick overview of the results. Having divided the result set into clusters, the user can narrow down his search further by selecting a cluster. This resembles query refinement but avoids the need to query the search engine for each step. Evaluations done using the grouper system indicate that users tend to investigate more documents per query than in normal search engines. It is assumed that this is because the user clicks on the desired cluster rather than reformulating his query. The evaluation also indicates that once one interesting document has been found, users often find other interesting documents in the same cluster.

The majority of the current search engines generate a huge list in reply to a user query. This result is normally ranked by using ranking criteria such as page rank or relevancy to the query. However, this list is extremely inconvenient to users, since it expects them to look into each page sequentially in an exhaustive manner to find the relevant information. As a result, most users only search for an initial few Web pages on the list. Thus many other relevant information can be overlooked.

The clustering method [19] is one such solution to overcome this problem. Instead of a sequential list, it groups the search results into clusters and labels these with representative words for each cluster. These labeled clusters of search results are exposed to users. The clustering method provides benefits in terms of reduced size of information provided to the end users. The clusters of items with common semantic and/or other characteristics can guide users in refining their original queries, to zoom in on smaller clusters, and drill down through subgroups. Search result clustering has several specific requirements that may not be essential for other cluster algorithms.

First, search result clustering should allow fast clustering and rapid generation of a label on the fly, since it is an online process. This requirement can be met by adopting “snippets” rather than entire documents of a search result set. Second, labels annotated for clusters should be meaningful to users because they are presented to users as a general view of results. For this reason, recent search result clustering research focuses on selecting meaningful labels. This differs from general clustering which focuses on the similarity of documents.

The Lingo algorithm proposed uses frequent phrases to identify candidate cluster labels and then assigns snippets to these lables. The extension of this lingo algorithm by adding semantic recognition to the frequent extraction phase is presented in [1].

In this study, a collaborative proximity-based fuzzy clustering [27] is used to discover a structure of web information by a prudent reliance on the structures in the spaces of semantics and data. The method focuses on the reconciliation between the two separated facets of web information and a combination of results leading to a comprehensive data organization. The information arranged in this manner can provide an integral description of web resources. This style of processing is explicitly implied by the findings as to the relevance of the distinction made with regard to these two spaces. This approach dwells on some existing mechanisms of fuzzy clustering in particular fuzzy C-means to complete a thorough arrangement of collections of Semantic Web documents (SWDs)[14], according to their facet-based characteristics. Through the proposed collaborative clustering Fig.13, a collection of homogeneous clusters can be built. Given these constructs, to look at clusters of web resources which are useful to formulate the query and to drive a search toward some “similar” documents existing on the web.

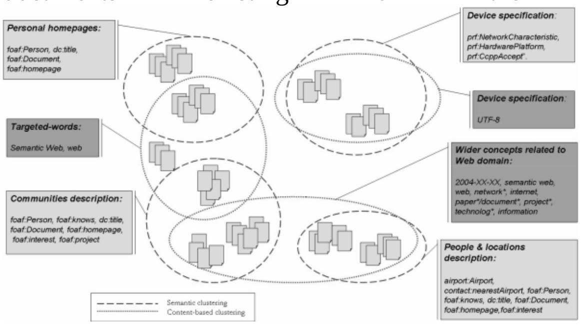

Personal homepages:
foaf:Person, dc:title, foaf:Document, foaf:homepage
Targeted-words:
Semantic Web, web
Communities description:
foaf:Person, foaf:knows, dc:title, foaf:Document, foaf:homepage, foaf:interest, foaf:project
Device specification:
prf:NetworkCharacteristic, prf:HardwarePlatform, prf:CcppAccept*.
Device specification:
UTF-8
Wider concepts related to Web domain:
2004-XX-XX, semantic web, web, network*, internet, paper*/document*, project*, technolog*, information
People & locations description:
airport:Airport, contact:nearestAirport, foaf:Person, foaf:knows, dc:title, foaf:Document, foaf:homepage, foaf:interest
Semantic clustering
Content-based clustering

Fig. 13. Semantic and content-based clustering.

3.5 Crawler Based Search

“Swoogle” [14] is a crawler-based indexing and retrieval system for the Semantic Web. It extracts metadata for each discovered document, and computes relations between documents. Discovered documents are also indexed by an information retrieval system to find relevant documents and to compute the similarity among a set of documents. One of the interesting properties is computing *ontology rank*, a measure of the importance of a Semantic Web document. As shown in Fig.14, Swoogle's architecture can be broken into four major components: SWD discovery, metadata creation, data analysis, and interface. This architecture is data centric and extensible; components work independently and interact with one another through a database.

The ***SWD discovery*** component discovers potential SWDs throughout the Web and keeps up-to-date information about SWDs.

The ***metadata creation*** component caches a snapshot of a SWD and generates objective metadata about SWDs at both the syntax level and the semantic level.

The ***data analysis*** component uses the cached SWDs and the created metadata to derive analytical reports, such as classification of SWOs and SWDBs, rank of SWDs, and the IR index of SWDs.

The ***interface*** component focuses on providing data services to the Semantic Web community.

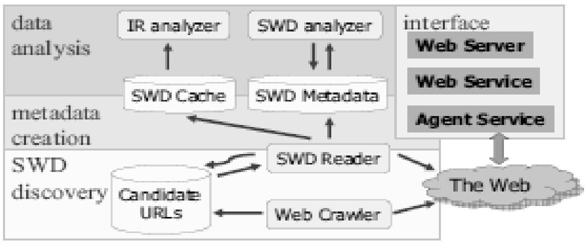


Fig 14.The architecture of Swoogle

Swoogle is improved by adding user preferences and interests to provide user a set of personalized results. In this paper the author proposes, architecture for a Personalized Semantic Search Engine (PSSE) [21]. PSSE is a crawler-based search engine that makes use of multi-crawlers to collect resources from both semantic as well as traditional web resources. In order to reduce processing time, web pages' graph is clustered, then clusters are annotated using document annotation agents that work in parallel. Annotation agents use methods of ontology matching to find resources of the semantic web as well as information extraction techniques. System ranks resources based on a final score that's calculated based on traditional link analysis, content analysis and a weighted user profile for more personalized results.

PSSE Architecture: As Fig.15 depicts, the processes of PSSE are separated into an offline and an online part. The offline part includes crawling and preprocessing processes. The online phase includes query processing and result ranking.

Offline Phase In this phase, crawling the World Wide Web and preprocessing of crawled pages take place.

Crawler PSSE uses Multi-crawlers (web spiders) that traverse World Wide Web, collect web resources and store them in database. Crawlers work with the aid of information extraction techniques to find link information in the retrieved pages.

Preprocessor The preprocessor is used to maintain resources that are downloaded from Web sites. The main task of query Indexer and link analyzer is to cluster the crawled web documents to enable parallel processing. This can be done in three steps: first indexer and link analyzer builds a graph of the crawled pages. Link analysis is then performed to calculate authoritativeness of web pages. And finally the graph is clustered by identifying its connected components. These clusters are then annotated by annotation agents that work in parallel to reduce processing time. Afterwards, annotations are weighted so as to determine their relevancy to web resource using term relevancy evaluator.

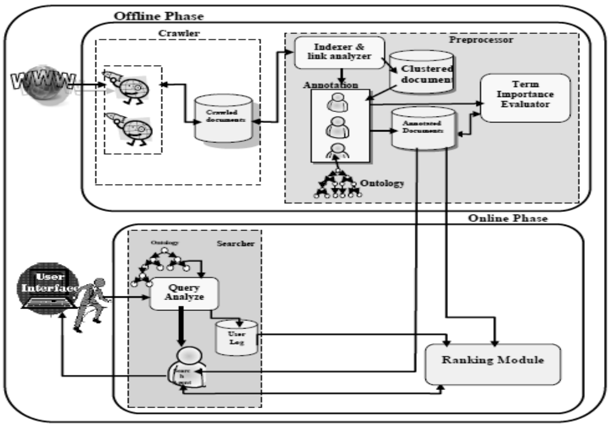


Fig 15. Architecture of PSSE

3.6 Other Directions

An approach based on metadata is discussed by Fabio Silva et al [8]. This work proposes a model to find information items with similar semantic content that a given user's query. The information items internal representation is based on user interest groups, called "semantic cases". The model also defines a similarity measure for ordering the results based on semantic distance between semantic cases items.

An annotation process extracts the metadata which is used to build the internal representation of documents and queries. Finally the matching process that uses concepts is used to find related documents and a semantic similarity function for the retrieval results ranking (Fig. 16). The main limitation of this model is the incompleteness of the conceptualization. The annotation process must be supported by ontology learning, to discover new items.

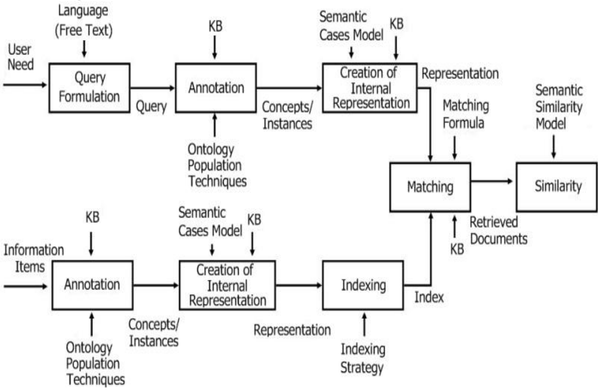


Fig 16. Overview of the proposed information retrieval process

The architecture for Developing a semantic-enable information retrieval mechanism [16] handles the processing, recognition, extraction, extensions and matching of content semantics to achieve the following objectives. i. Analyse and determine the semantic features of content, to develop a semantic pattern that represents semantic features of the content, and to structuralize and materialize semantic features; ii. Analyse user's query and extend its implied semantics through semantic extension so as to identify more semantic features for matching iii) Generate contents with approximate semantics by matching against the extended query to provide correct contents to the queriest.

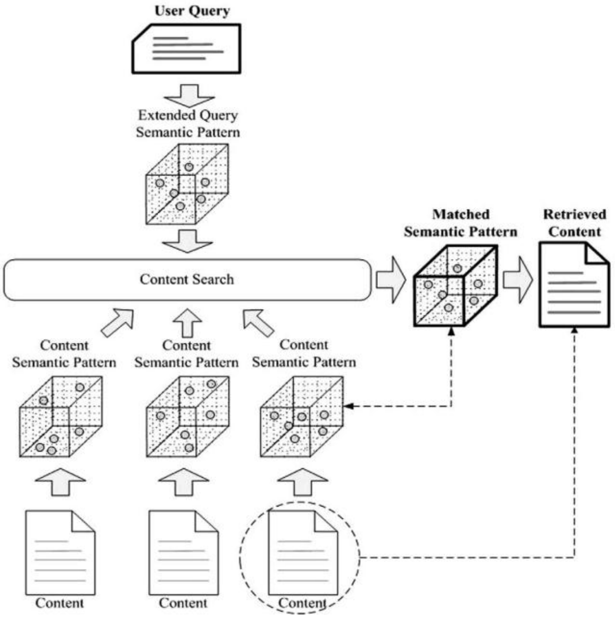


Fig. 17. Scenario of semantic-enable information retrieval

This architecture contains the core technologies such as Semantic determination and extraction, Semantic extension, Semantic pattern clustering and matching. In addition to semantic-based information retrieval, the proposed system has two main features: i) Latent semantic analysis to generate more semantics for matching, thereby solving the problem of insufficient information for query; ii) Semantic clustering model which identifies the corresponding document category for the query and then performs content matching in that category thereby improving matching accuracy.

An overview of Semantic Search Systems [29] discussed with the help of a framework which has six components responsible for data acquisition, knowledge acquisition, data integration and consolidations, semantic search mechanisms, semantic search services , and result presentation.

The ***Semantic data acquisition*** will provide different solutions to collect all the structured, semi structured and unstructured data. The collected data is transformed into structured data using ***Knowledge acquisition*** component.

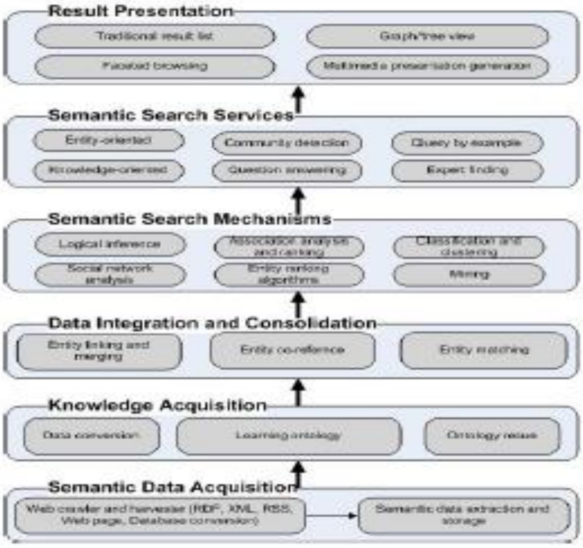


Fig 18. A Semantic Search Framework

The ***Data integration and consolidation*** component summarises solutions for a problem arisen from the previous stage. ***Search mechanisms*** component deals with various techniques based on which semantic search services are implemented. ***Semantic Search Services*** provide an abstract model of the functionalities a semantic search engine offers. Finally the ***Result Presentation*** component presents the search results to the user.

4. Conclusion

In this paper, we survey the various architectures of Semantic Search Systems and classify them in five dimensions: Relation Based, Ontology Based, Similarity Based, Cluster Based, Crawler Based Search systems. The first dimension process the user's query based on the keyword- concept pair. The second dimension will find the relevant Ontology to the user. The third criteria ranks the Ontologies based on the rank calculated and arranged, the most relevant ontology is submitted to the user. In the fourth criteria, instead of linear list the results are presented in the form of clusters with appropriate labels. The last dimension makes use of the crawler to collect the semantic documents and to find the relevant information on the retrieved paper.

There are several points to make from this survey as a future direction. First the semantic searching mainly focused on the trust and the quality of knowledge which varies largely from source to source. Effective ranking algorithms are needed to distill most trustworthy and quality information. The second aspect is ontology-based research focused mainly on integrity of the Domain Ontology, Automatic Ontology Evolution and Ontology Learning. Since the web is decentralized and heterogeneous, even on the same domain it seems impossible for all web pages to use the same ontology. So study on semantic interoperability will be needed. The third aspect is assigning weights relies on the user explicitly assigning numerical weights to properties through the query interface and hence imposes some overhead to the users. Methods should be explored to assign weights automatically through relevance feedback strategy and predicting users preference. Another promising direction is to incorporate rules to support more powerful reasoning based intelligent semantic search.

References

[1] Ahmed Samesh, Amar Kadray ,"Semantic Web Search Results Clustering Using Lingo and WordNet" , International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 1,No.2, June 2010.

[2] Aleman-Meza B, Halaschek C, Arpinar IB, Sheth "A Context-aware semantic association ranking", Semantic web and databases workshop proceedings. Berlin, Germany, September pp.7–8, 2003.

[3] Amit Pisharody, Howard E.Michel,"A Search Engine Technique Using Relation Based Keywords", Proceedings of the 2005 International Conference on Artificial Intelligence, IC-AI 05.

[4] Baeza-Yates R A and B A Riberiro-Neto, "Modern Information Retrieval",ACM Press/Addision-Wesley,1999.

[5] Broekstra J, EhrigM, Haase P, van Harmelen F,KampmanA, Sabou M,et al. "A metadata model for semantics-based peer-to-peer systems", Proceedings of the WWW'03 workshop on semantics in peer-to-peer and grid computing. Budapest, Hungary, pp.20–24, May 2003.

[6] Burners-Lee T, J Hendler and O Lassila, " The Semantic Web", Scientific American, Vol.284,No.4.2001.

[7] Darijus Strasunskas and Stein Tomassen "On Variety of Semantic Search Systems and Their Evaluation Methods" International Conference on nformation Management and Evaluation (ICIME 2010), pp. 380-387, Mar 2010.

[8] Fabio Silva, Rosario Girardi, and Lucas Drumond "An Information Retrieval Model for the Semantic Web" 2009 Sixth International Conference on Information Technology: New Generations,2009.

[9] Fabrizio Lamberti, Andrea Sanna, Claudio Demartini, " A Relation-Based Page Rank Algorithm for Semantic Web Search Engines " IEEE Trans. Knowledge and Data Eng., vol. 21, No.1, pp. 123-135, Jan. 2009.

[10] Gopinath Ganapathy1 and S. Sagayaraj "Studies on Architectural Aspects of Searching using Semantic Technologies" International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 1, No. 2,pp. 119-126, June 2010.

[11] Guha.R, McCool.R and Miller.E, "Semantic Search", International Conference on World Wide Web, pp.700-709,2003.

[12] Guobing Zou Bofeng Zhang Yanglan Gan Jianwen Zhang "An Ontology-based Methodology for Semantic Expansion Search", Fifth International Conference on Fuzzy Systems and Knowledge Discovery 2008.

[13] Lee.J, Kim.M, and Lee.Y, "Information Retrieval Based on Conceptual Distance in IS-A Hierarchies," Documentation, vol 49. pp.188-207,1993.

[14] Li Ding, , Finin.L, , Joshi.T.W, , Pan.A, Scott Cost.R, Peng.R, Reddivari.Y Doshi.P, Sachs.S "Swoogle: a search and metadata engine for the semantic web", CIKM 2004, pp.652-659,2004.

[15] Lixin Hana,b,c, Linping Suna, Guihai Chenb, Li Xieb "ADSS: An approach to determining semantic similarity Advances in Engineering Software" Elsevier, Science Direct 37, pp. 129–132,(2006).

[16] Ming-Yen Chen, Hui-Chuan Chu, Yuh-Min Chen "Developing a semantic-enable information retrieval mechanism" Elsevier, Expert Systems with Applications 37 pp. 322–340,(2010).

[17] Pablo Castells, Miriam Ferna´ndez, and David Vallet " An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval" IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 2, pp. 261-272, February 2007.

[18] Pekar V, Staab S. "Word classification based on combined measures of distributional and semantic similarity", Proceedings of the research note sessions of the 10th conference of the European chapter of the association for computational linguistics (EACL'03). Budapest,Hungary, pp. 12–17 ,April 2003.

[19] Ramesh Singh, Dhruv Dhingra, and Aman Arora "SSCHISM—A Web search engine using semantic taxonomy "IEEE Potentials, pp 36-40,2010.

[20] Ravi Sankar V, A.Damodaram and P.Radha Krishna "Ranking Ontologies Based on OWL Language Constructs", Information Technology Journal 9(3):,ISSN 1812-5638 pp. 553-560,2010.

[21] Riad A.M., Hamdy.K Elminir., Mohamed Abu ElSoud, Sahar. F. Sabbeh. "PSSE: An Architecture For A Personalized Semantic Search Engine" ,International Journal on Advances in Information Sciences and Service Sciences Volume 2,Number 1, pp. 102-112 ,March 2010.

[22] Rodriguez M, Egenhofer M. "Determining semantic similarity among entity classes from different ontologies", IEEE Transactions on Knowledge and Data Engineering 15(2), pp. 442–56,2003.

[23] Rudi L. Cilibrasi and Paul M.B. Vita´ nyi "The Google Similarity Distance" IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 3, pp.370-383, March 2007.

[24] Song Jun-feng, Zhang Wei-ming, Xiao Wei-dong, Li Guo-hui, Xu Zhen-ning "Ontology-Based Information Retrieval Model for the Semantic Web "International Symposium on Intelligent Information Technology Application Workshops 2008.

[25] Sridevi.U.K, Nagaveni "Ontology based Similarity Measure in Document Ranking", International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 26 pp.135-139,2010.

[26] Thomas Bernecker, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, and Andreas Zuefle "Scalable Probabilistic Similarity Ranking in Uncertain Databases" IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 9, pp.1234-1246, September 2010.

[27] Vincenzo Loia, Witold Pedrycz, and Sabrina Senatore"Semantic Web Content Analysis: A Study in Proximity-Based Collaborative Clustering" IEEE Transactions on Fuzzy Systems, Vol.15, No. 6, pp.1294-1312 ,December 2007.

[28] Wei-Dong Fang, Ling Zhang, Yan Xuan Wang, Shou-Bin Dong " Toward a Semantic Seach Engine Based On Ontologies" Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, pp.18-21, August 2005.

[29] Wei Wang, Payam M. Barnaghi, Andrzej Bargiela, "Search with Meanings:An Overview of Semantic Search Systems", International journal of Communications of SIWN, Vol. 3, pp. 76-82, June 2008.

[30] Xiaomeng Su , Jon Atle Gulla "An information retrieval approach to ontology mapping Data & Knowledge Engineering" Elsevier, ScienceDirect 58 pp. 47–69 ,(2006) .

[31] Yi Jin, Zhuying Lin, Hongwei Lin" The Research of Search Engine Based on Semantic Web " International Symposium on Intelligent Information Technology Application Workshops,2008.

[32] Yi Zhang, Wamberto Vasconcelos, Derek Sleeman "OntoSearch:An Ontology Search Engine " Research and Development in Intelligent Systems XXI 2005, Session 1a, DOI: 10.1007/1-84628-102-4_5,pp. 58-69,2005.

[33] Yuangui Lei, Victoria Uren, Enrico Motta, "SemSearch: A Search Engine for the Semantic Web", EKAW 2006, Springer, pp:238-245, 2006.

[34] Yufei Li, Yuan Wang, and Xiaotao Huang "A Relation - Based Search Engine in Semantic Web," IEEE Trans. Knowledge and Data Eng., vol. 19, No.2,pp. 273-282, Feb. 2007.

[35] Zheng Gu, Song-Nian Yu, "Ontology-Based Inverted Tables in Information Retrieval System", Third International conference on Semantics,Knowledge and Grid 2007.

[36] http://www.w3.org/RDF/, Resource Description Framework (RDF); World Wide Web Consortium; August, 2003.

Author Biographies

Muthuram Thangaraj received his post-graduate degree in computer science from Alagappa University, Karaikudi, M.Tech. degree in Computer Science from Pondicherry University and Ph.D. degree in Computer Science from Madurai Kamaraj University, Madurai,TN,South India in 2006. He is now the Associate Professor of Computer Science Department at M.K.University. He is an active researcher in Webmining, Semantic Web and Inforamtion Retrieval and has published more than 35 papers in Journals and Conference Proceedings.

Sujatha G was born on 14.7.71 in Madurai. She received her M.C.A. degree in 1994 from Alagappa University, Karaikudi,M.Phil. in Computer Science from Mother Teresa University, Kodaikanal in 2000.Currently she is an Assistant Professor in Computer Science, Sri Meenakshi Govt. College for Women, Madurai, Tamil Nadu, India. Her research interests are focused on Semantic Search and the Information Retrieval. sujisekar05@rediffmail.com.

Bridging the Gap between UML and Hardware Description Languages at Early Stages of Embedded Systems Development

Fateh Boutekkouk[1]

[1]Department of computer science, Larbi Ben M'hedi University,
BP 358, Route de Constantine, Oum El Bouaghi, 04000, Algeria
fateh_boutekkouk@yahoo.fr

***Abstract*:** This work deals with automatic Hardware Description Languages (HDLs) code generation from UML 2.0 models at early stages of embedded systems development. In our case, we target two standard HDLs which are SystemC and VHDL. A particularity of our proposed approach is the fact that HDLs code generation process is performed through two levels of abstraction. In the first level, we use UML hierarchic sequence diagrams to generate a HDL code that targets algorithmic space exploration and simulation eventually. In the second level of abstraction, messages that occur in sequence diagrams are implemented using UML activity diagrams whose state actions are expressed in the C++ Action Language included in the Rhapsody environment from which a full HDL code is generated for both simulation and synthesis. We have developed two macros for SystemC and VHDL code generation and integrated them as tool boxes in the Rhapsdoy environment.

***Keywords*:** Embedded Systems, UML, SystemC, VHDL, Simulation, Synthesis**.**

1. Introduction

We can define Embedded Systems (ESs) [9] as application-specific computers, masquerading as non-computers that interact with the physical world and must perform a small set of tasks cheaply and efficiently. ESs have specific characteristics such as heterogeneity (hardware / software), ability to react, criticality, real time, and consumption constraints.

Modern ESs are able to execute very complex algorithms ranging from control, telecommunication to media high performance applications implemented in only one chip (SOC: System-On-a-Chip) [10].

The ever complexity of embedded systems (ESs) design has pushed researchers in the field to raise the level of abstraction and exploit recent Software Engineering technologies such as object technology and in particular the Unified Modeling Language (UML) [6].

ESs designers are now confronted with the challenge of how to close the gap between UML and the well practiced Hardware Description Language (HDL) in ESs world such as SystemC [20] and VHDL [23].

Since UML was originally introduced in the software field, most commercial tools generate software code such as C, C++, and Java from UML models. However, there is a lack of tools that can synthesize UML models into HDL descriptions.

Our objective is to raise the level from which HDL descriptions can be generated to perform quick algorithmic space exploration, simulation and synthesis eventually. Thus a refinement directed approach seems inevitable to bridge the gap smoothly between UML models and HDLs descriptions.

To address this problem, we have proposed a flow that permits automatic HDL code generation from UML models at two levels of abstraction. The first level corresponds to HDL code generation from UML sequence diagrams without implementing messages. Thus the code generated at this stage is oriented to algorithmic space exploration and simulation eventually since the obtained code consists only of processes input/output ports, processes sensitivity lists, dependencies between processes, and signals. The second level of abstraction is viewed as a refinement of the first level where messages are implemented using UML activity diagrams whose actions are expressed in the C++ Action Language included in the Rhapsody environment [15]. At this stage, the generated code is dedicated to both simulation and synthesis. In this paper, our main contribution is the development of a tool that can generate SystemC and VHDL code from UML models following a refinement directed approach. The rest of this paper is organized as follows: section two is dedicated to related works concerning the synthesis of UML models to SystemC and VHDL code. Section three gives an overview of VHDL and SystemC languages. Our proposed flow with an illustrative example is discussed in section four. The implementation of our tool and a case study is discussed in section five before concluding.

2. Related Work

In this section, we try to present briefly some pertinent woks targeting the generation of VHDL and SystemC codes from UML models.

The authors in [9] proposed the synthesis of state diagrams into VHDL.

In [12], the authors presented a technique for generating VHDL descriptions from a subset of UML, and a set of rules to transform UML classes and Statecharts to VHDL.

The authors in [4] and [5] used SMDL (the language with formal semantics and high-level concepts such as states, queues and events) as an intermediary language to generate VHDL code from UML Statecharts and activity diagrams.

A Model Driven Architecture (MDA) approach for generating VHDL code from UML models was proposed in [1], [8], and [17]. In [8], the authors used UML Meta-model to generate different platform specific implementations.

In [17], the authors defined a set of rules to map UML to VHDL in a practical code generator.

In [16], the authors presented a UML/SystemC profile for SystemC code generation from UML structural and Statecharts diagrams.

In [21], the authors developed a tool for UML synthesis called: *Chip Fryer* that can generate VHDL code from XMI representation of UML models. The input model consists of class, object diagrams, and state machines. Actions are expressed in a C++ action language.

In [24], the authors proposed a UML/MDA approach called *MoPCoM* methodology that permits automatic VHDL and SystemC code generation from UML models and MARTE profile by means of MDA techniques. Input models are focused on UML class, component, and Statecharts diagrams.

Contrary to these works, our approach tries to generate VHDL and SystemC codes automatically at early stages of ESs development from UML sequence diagrams in a first step then from UML activity diagrams in a second step.

3. VHDL and SystemC

3.1 VHDL

VHDL (VHSIC Hardware Description Language) [2], [3], [23] is an industrial standard HDL. It looks similar to programming language ADA and used for both simulation and synthesis.

Now VHDL is governed by IEEE standards and very popular for European design houses. VHDL models consist of an external part (entity) that defines the Inputs/Outputs of the model and the internal part that describes the operation of the model (the architecture). The Entity declaration format looks like:

```
entity entity_name is
port (signal_name(s): mode signal_type;
:
signal_name(s): mode signal_type);
end entity entity_name;
```

mode describes the direction of transferred data through port (*in*, *out*, or *inout*); *signal_type* defines the signal(s) type.

The Architecture format looks like:

```
architecture architecture_name of entity_name is
begin
:
end architecture architecture_name;
```

VHDL designs can be written in three different styles: structural, data flow, and behavioural. Of course, these three styles can be mixed. Structural descriptions describe the interconnection of hierarchy and are useful for designs reuse. They consist of component instantiation statements (i.e. *port map* instruction) which are concurrent statements.

Behavioural descriptions are focused on the *process* concept. The latter is used in two ways:

For combinational logic, we mention the list of all process input signals after the keyword *process*. The general form is:

```
process (signal_names)
begin
.....
end process;
```

For sequential logic, two cases occur:

In the first case, the sensitivity list is empty, but statements inside the process must include wait statements;

In the second case, the sensitivity list contains the clock signal and the statements are within an *if* statement.

The general form is as follows:

```
process (clock)
begin
if clock and clock'event then
....
end if;
end process;
```

Processes communicate via signals. Many processes can be put in one architecture. VHDL supports classical language data types such as: *boolean*, *character*, *integer*, *real*, and *string* and control statements such as *if*, *loop*, and *case*. In addition, VHDL has the types: *bit*, *bit_vector*, and the IEEE 1164-standard-logic types that are *std_logic* and *std_logic_vector*. For more details on VHDL, one can refer to [23].

3.2 SystemC

SystemC [18], [19], [20] is an extension of C++ language for SOC modeling and simulation. Various versions of the language have appeared but we consider SystemC2.0. SystemC structural designs are focused on modules. A module contains ports, interfaces, channels, processes, and eventually other modules. In SystemC, concurrent behaviors are modeled using processes. A process has a sensitivity list that includes the set of signals to which it is sensitive. This list can be either static (pre-specified before simulation starts) or dynamic.

SystemC processes execute concurrently and may suspend on *wait()* statements. Such processes requiring their own independent execution stack are called “SC_THREADs”. When the only signal triggering a process is the clock signal ‘*clk*’ we obtain what we call “SC_CTHREAD” (clocked thread process). Certain processes do not actually require an independent execution stack and cannot suspended on *wait()* statement. Such processes are termed “SC_METHODs”. SC_METHOD processes execute in zero simulation time and returns control back to the simulation kernel.

The following code [19] presents a SystemC module named *display* with an input port *din*, and an SC_METHOD called *print_data* which is sensible to *din*. For each SystemC module there are two files: *.h* for ports, functions, variables, and processes declaration and *.cc* for process and functions implementation. *systemc.h* designates the SystemC library file.

```
// display.h
#include "systemc.h"
#include "packet.h"

SC_MODULE(display) {
sc_in<long> din; // input port
void print_data();
// Constructor
SC_CTOR(display) {
SC_METHOD(print_data); // Method process to print data
sensitive << din;
}
};
// display.cc
#include "display.h"
void display::print_data() {
cout <<"Display:Data Value Received, Data = "<< din <<
"\n";
```

4. Our proposed flow

As showed in figure 1, our proposed flow starts by capturing system requirements as a set of related uses cases and actors. At this stage, we use UML uses cases with *'include'* and *'extend'* relations. Figure 2 gives an example of modelling with use cases diagram. In this example, we have one actor and two use cases named *usecase_0* and *usecase_1*. *usecase_0* is related to *usecase_1* by the 'include' relation.

Each use case diagram is then refined into a set of interacting objects showing a possible scenario. At this stage, we use UML sequence diagram. The 'include' relation is modelled as an unconditional call of the use case child while the 'extend' relation is an optional call subject to some condition. Figure 3 shows a possible implementation of use cases using hierarchic sequence diagrams. In this example, we model *usecase_0* as the parent use case using sequence diagram with three interacting objects (class's instances) *class_0*, *class_1*, and *class_2* and an external object that represents the environment (*Env*). *usecase_1* is modelled as a child sequence diagram invoking by a call from the environment. In order to model the 'extend' relation, we add a conditional call invoking the child sequence diagram (*usecase*_2 in figure 4). From UML sequence diagrams, VHDL and SystemC codes are generated automatically using the VB API which is integrated in the Rhapsody environment. This API offers the necessary functions and commands that permit the manipulation of UML diagrams and then the extraction of information needed for HDL code generation as text files. The generated code in this step will be used for algorithmic space exploration and simulation eventually.

We have used three techniques for HDL code generation process:

1st technique: each message is considered as a VHDL process/SystemC SC_METHOD.

2nd technique: each end-to-end scenario is considered as a VHDL process/ SystemC SC_THREAD.

3rd technique: each object is considered as a VHDL process/ SystemC SC_THREAD.

For each technique, two styles of VHDL descriptions are generated: structural using VHDL mapping instructions and behavioural using the VHDL process concept. Dashed lines in figure 2 enable the designer to modify his/her design according to simulation results. VHDL/SystemC simulation and/or synthesis are performed using available commercial tools such as *ModelSim* (ModelSim) or *SystemC* simulator.

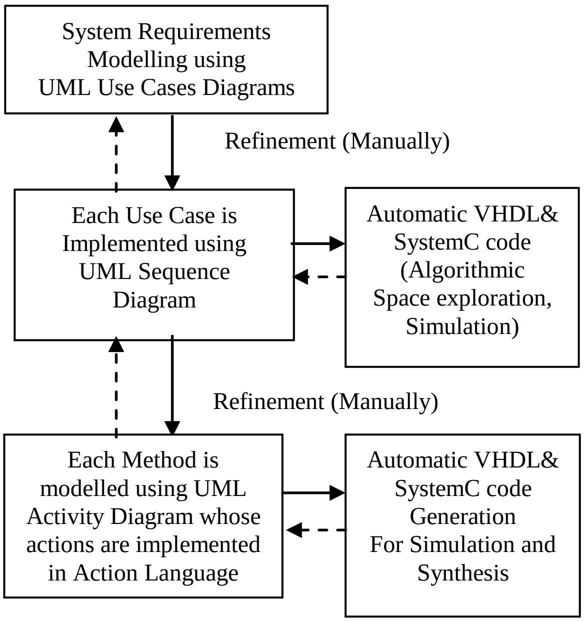

System Requirements Modelling using UML Use Cases Diagrams
Refinement (Manually)
Each Use Case is Implemented using UML Sequence Diagram
Automatic VHDL& SystemC code (Algorithmic Space exploration, Simulation)
Refinement (Manually)
Each Method is modelled using UML Activity Diagram whose actions are implemented in Action Language
Automatic VHDL& SystemC code Generation For Simulation and Synthesis

Figure 1. Our proposed flow

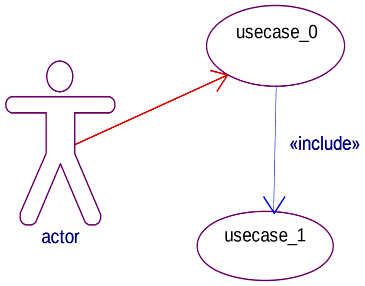


Figure 2. Example of UML use cases diagram

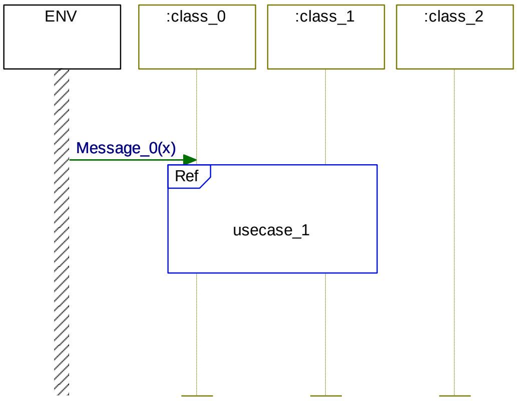


Figure 3. Possible implementation of 'include' relation

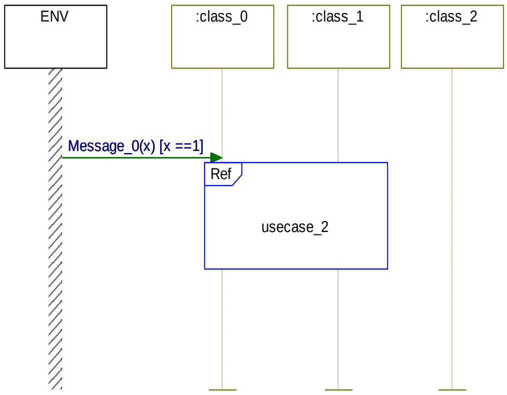


Figure 4. Possible implementation of ‘extend’ relation

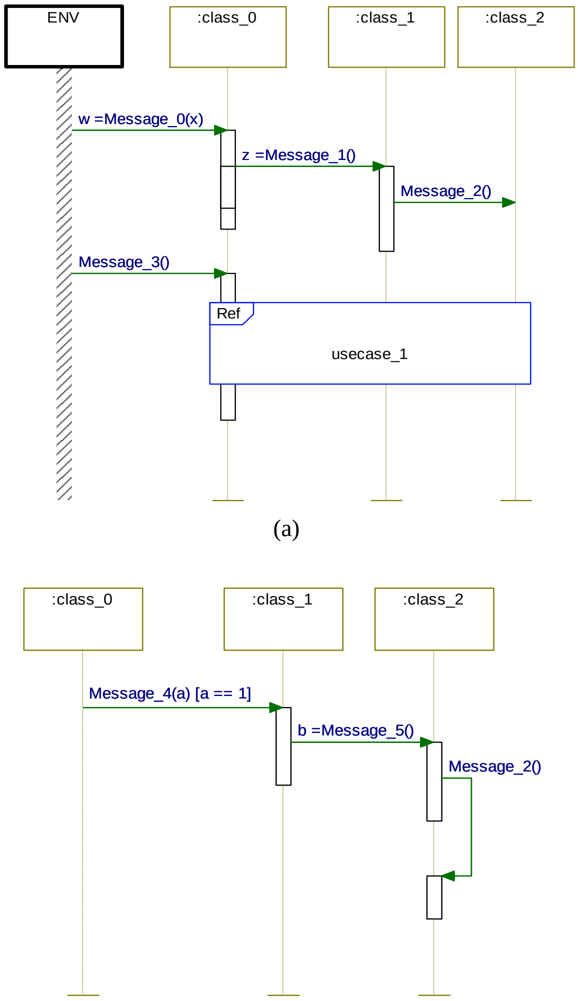
4.1 Illustrative example

In order to motivate our proposed approach, we try to apply the HDL code generation process on an example whose use case diagram is illustrated in figure 2. In this example, we assume that we have an actor and two use cases named *usecase_0* and *usecase_1* that are related by an ‘include’ relation. Both *usecase_0* and *usecase_1* are implemented using UML sequence diagrams as showed in figure 5. In the following sections, we try to explain the three techniques for VHDL/SystemC code generation from UML sequence diagrams.

4.2 First technique

In this technique, each message is mapped to a VHDL process or a SystemC SC_METHOD.

Methods arguments are transformed to input ports while returned values are mapped to output ports. To each call to a message, we add a Boolean input port that corresponds to the event to which process is sensible and a Boolean output port that corresponds to control return. From figure 5, we observe that *message_2* is used in both *usecase_0* and *usecase_1*. Such a common message will be mapped to a SC_METHOD process in a separate module. Two styles of VHDL descriptions are generated: the behavioural description and structural description. In the former, all generated processes from children sequence diagrams are put in one architecture that corresponds to the main sequence diagram. In the latter, we consider children sequence diagrams as sub entities reflecting the hierarchy of the design. Table 1 shows the correspondence between UML and VHDL/SystemC concepts.



(a)

(b)

Figure 5. Example of hierarchic sequence diagrams (a) parent sequence diagram (usecase_0); (b) child sequence diagram (usecase_1)

Assume that we have a message with two integer arguments (*a* and *b*) and an integer return value (*x*): *x = message(a,b)*. The corresponding VHDL code for this message is as follows:

```
message : process is
variable arg1, arg2, result : integer;
begin
wait until cal = true;    -- cal read
cal <= false;             -- cal write
arg1 := a;
arg2 := b;
-- message body
x <= result;              -- x write
ret <= true;              -- ret write
end process message;
```

Table 1. Correspondence between UML and VHDL/SystemC for the first technique

UML concept	VHDL (behavioral /structural)	SystemC
Message	Process/Entity	SC_METHOD
Common message	Process/Entity	SC_METHOD in a separate module
Argument	in port	sc_in <type> port
Returned value	out port	sc_out <type> port
call	inout port (boolean)	sc_inout <bool> port
Control return	out port (boolean)	sc_out <bool> port
Child sequence diagram	sub entity (structural)	sub module
Top level model	Test bench	sc_main()

arg1 and *arg2* are two variables used to stock the two arguments coming from the two ports (signals) *a* and *b*.
result is a variable used to stock the returned value in the port *x*. We use the Boolean ports *cal* and *ret* to specify the message invoking and the return of the control to the caller respectively. The meaning of this VHDL code is as follows:
The process *message* will be blocked until the occurrence of the signal *cal* (*cal* = true). After that, the process resumes its execution: sets *cal* to false; stock the arguments coming from the input ports *a* and *b* into variables *arg1* and *arg2*; performs some computations; stocks the result of computation into output port *x*; sets the signal *ret* to true.
Similarly, The VHDL code for the caller looks like:

```
caller : process is
variable val : integer;
begin
-- instructions
cal <= true;              -- cal write
a <= " ";                 -- initialization
b <= " ";
wait until ret = true;      -- ret read
ret <= false;             -- ret write
val:= x;                  -- x read
-- Remaining instructions
end process caller;
```

The meaning of this VHDL code is as follows:
After performing some computations, the process *caller* sets the signal *cal* to true; initializes the arguments ports *a* and *b*; blocked until the occurrence of the signal *ret* (*ret* = true). After that, the process resumes its execution: sets *ret* to false; stocks the content of port *x* into variable *val*; performs the remaining computation.
The corresponding SystemC code for this message is as follows:

```
// module1.h
# include "systemc.h"
SC_MODULE(module1){
sc_in<int> a;
sc_in<int> b;
sc_out<int> x;
sc_inout<bool> cal;
sc_out<bool> ret;
void message();
SC_CTOR(module1) {
SC_METHOD(message);
sensitive << cal; }};
// module1.cc
#include "module1.h"
void module1::message() {
int var1, var2, result;
while cal == 0 ;
cal = 0;     // cal = false;
var1 = a;
var2 = b;
// message body
x = result;
ret = 1; }          //  ret = true;
```

SC_METHOD message is sensitive to the signal *cal*.
The SystemC code for the caller is as follows:

```
// module2.h
# include "systemc.h"
SC_MODULE(module2){
sc_in<int> x;
sc_inout<bool> ret;
sc_out<int> a;
sc_out<int> b;
sc_out<bool> cal;
void caller();
SC_CTOR(module2) {
SC_METHOD(caller);
sensitive << ****; // some ports
}};
// module2.cc
#include "module2.h"
void module2::caller() {
int result;
// instructions;
cal = 1;     // cal = true;
a = " ";  // arguments initialization
b = " ";
While ret == 0 ;
ret = 0;
result = x;
// remaining instructions
}
```

Note that SC_METHOD processes *message* and *caller* are put in two distinct modules: *module1* and *module2* respectively. However, if we put them into one module, all ports become sc*_inout*.
By applying this technique on our example, we obtain six (6) VHDL processes and six SC_METHOD processes that are: *Message_0*, *Message_1*, *Message_2*, *Message_3*, *Message_4*, and *Message_5*. In the VHDL behavioural style, all processes are put in one architecture. The entity includes all processes ports. Assume that all messages arguments and return values are integers. *cal0*, *cal1*, *cal2*, *cal3*, *cal4*, and

cal5 designate Boolean ports for *message_0*, *message_1*, *message_2*, *message_3 message_4*, and *message_5* calls respectively. *arg0* and *arg4* designate ports for *message_0* and *message_4* arguments respectively. *val0*, *val1*, and *val5* designate ports for *message_0*, *message_1*, and *message_5* returned values respectively. *ret0*, *ret1*, *ret2*, *ret3*, *ret4*, and *ret5* designate Boolean ports for messages controls return.

The corresponding VHDL code for the behavioural description is as follows:

```
entity  usecase_0 is
port (cal0, cal1, cal2, cal3, cal4, cal5 : inout boolean; arg0 :
in integer; arg4 : inout integer; ret0, ret1, ret2, ret3, ret4,
ret5 : inout  boolean; val0: out integer; val1, val5 : inout
integer);
end entity usecase_0;

architecture system of usecase_0 is
begin
message_0 : process is
variable arg, val : integer;
begin
wait until cal0 = true;
cal0 <= false;
arg := arg0;
-- instructions
cal1 <= true ;
wait until ret1 = true ;
ret1 <= false ;
val := val1;
-- remaining instructions
val0 <= w;
ret0 <= true ;
end process message_0;

message_1 : process is
begin
wait until cal1 = true;
cal1 <= false;
-- instructions
cal2 <= true;
wait until ret2 = true;
ret2 <= false;
-- remaining instructions
val1 <= z;
ret1 <= true;
end process message_1;

message_2 : process is
begin
-- code
end process message_2;

message_3 : process is
variable temp : integer;
begin
wait until cal3 = true;
cal3 <= false;
-- instructions
if temp = 1 then
cal4 <= true;
arg4 <= temp;
wait until ret4 = true;
ret4 <= false;
end if
-- remaining instructions
ret3 <= true;
end process message_3;

message_4 : process is
-- code
end process message_4;

message_5 : process is
begin
-- code
end process message_5;
end architecture system;
```

The VHDL structural style is obtained by considering each process as a separate entity as well as all children sequence diagrams. For the sake of space, we do not show all messages VHDL code, rather than, we give the VHDL code only for *message_0*.

```
entity  message0 is
port (cal0 : inout boolean, cal1: out  boolean; ret0 : out
boolean, ret1: inout boolean; arg0 : in integer; val0 : out
integer; val1 : in integer);
end entity message0;

architecture basic of message0 is
begin
message_0 : process is
variable arg, val : integer;
begin
wait until cal0 = true;
cal0 <= false;
arg := arg0;
-- instructions
cal1 <= true ;
wait until ret1 = true ;
ret1 <= false ;
val:= val1;
-- remaining instructions
val0 <= w;
ret0 <= true ;
end process message_0;
end architecture basis;

entity  usecase_1 is
port (cal4 : inout boolean; arg4 : in integer; ret4 : out
boolean);
end entity usecase_1;

architecture struct of usecase_1 is
signal cal2, cal5, ret2, ret5 : boolean
signal val5 : integer;
begin
```

```
messag2 : entity work.message2(basic)
        port map (cal2,ret2);
messag4 : entity work.message4(basic)
        port map (cal4,cal5,ret4,ret5,arg4,val5);
messag5 : entity work.message5(basic)
        port map (cal5,cal2,ret5, ret2, val5);
end architecture struct;

architecture struct of usecase_0 is
signal ret1, cal1, cal2, ret2, cal4, ret4 : boolean;
signal arg4, val1 : integer;
begin
messag0 : entity work.message0(basic)
        port map (cal0, cal1, ret0, ret1,arg0, val0, val1);
messag1 : entity work.message1(basic)
        port map (cal1,ret2,ret1,cal2, val1);
messag2 : entity work.message2(basic)
        port map (cal2,ret2);
messag3 : entity work.message3(basic)
        port map (cal3, cal4, ret4,ret3, arg4);
usecase1: entity work.usecase_1(struct)
        port map (cal4,arg4,ret4);
end architecture struct;

entity  test_bench is
end entity test_bench;
architecture test_usecase_0 of test_bench is
signal cal0, cal3, ret0, ret3 : boolean;
signal arg0, val0 : integer;
begin
usecase0 : entity work.usecase_0(struct)
          port map(cal0, ret0, arg0, val0, cal3, ret3) ;
stimulus : process is
begin
cal0 <= true ;
ret0 <= false;
arg0 <= 500;
val0 <= 0;
cal3 <= true ;
ret3 <= true ;
end process stimulus;
end architecture test_usecase_0;
```

Since *message_2* is a common message, we put it in a separate module called *mess2*. Here, we have two modules: *usecase0* including SC_METHODS *message_0*, *message_*1, and *message_3*, and *usecase1*including *message_4*, and *message_5*.
The corresponding SystemC code is as follows:

```
// mess2.h
# include “systemc.h”
SC_MODULE(mess2){
sc_inout<bool> cal2;
sc_out<bool> ret2;
void message_2();
SC_CTOR(mess2) {
SC_METHOD(message_2);
sensitive << cal2;
}};
```

```
// mess2.cc
#include “mess2.h”
void mess2::message_2() {
while cal2 == 0 ;
cal2 = 0;
// message body;
ret2 = 1;}

// usecase1.h
# include “systemc.h”
SC_MODULE(usecase1){
sc_in<int> arg4;
sc_inout<int> val5;
sc_out<bool> cal2;
sc_inout<bool> ret2;
sc_inout<bool> cal4;
sc_inout<bool> cal5;
sc_inout<bool> ret5;
sc_out<bool> ret4;
void message_4();
void message_5();
SC_METHOD(message_4);
sensitive << cal4;
SC_METHOD(message_5);
sensitive << cal5;
}};

// usecase1.cc
void usecase1::message_4() {
int var, result;
while cal4 == 0;
cal4 = 0;
var = arg4;
// instructions
cal5 = 1;
while ret5 == 0;
ret5 = 0;
result = val5;
// remaining instructions
ret4 = 1;
}
void usecase1::message_5() {
// code
}
// usecase0.h
# include “systemc.h”
SC_MODULE(usecase0){
sc_in<int> arg0;
sc_inout<int> arg4;
sc_out<int> val0;
sc_inout<int> val1;
sc_inout<bool> cal0;
sc_inout<bool> cal1;
sc_out<bool> cal2;
sc_inout<bool> cal3;
sc_out<bool> cal4;
sc_out<bool> ret0;
sc_inout<bool> ret1;
```

```
sc_inout<bool> ret2;
sc_out<bool> ret3;
sc_inout<bool> ret4;
void message_0();
void message_1();
void message_3();
SC_CTOR(usecase0) {
SC_METHOD(message_0);
sensitive << cal0;
SC_METHOD(message_1);
sensitive << cal1;
SC_METHOD(message_3);
sensitive << cal3;
}};

// usecase0.cc
#include “usecase0.h”
void usecase0::message_0() {
 // code
};
void usecase1::message_1() {
// code
};
void usecase1::message_3() {
int var;
while cal3 == 0 ;
cal3 = 0;
// instructions
arg4 = var;
if arg4 = 1 {
cal4 = 1;
while ret4 == 0;
ret4 = 0;
}
// remaining instructions
ret3 = 1;
};
// main.cc
#include “mess2.h”
#include “usecase1.h”
#include “usecase0.h”
int  sc_main(int argc, char* argv[]) {
sc_signal<int> ARG0, ARG4, VAL0, VAL1;
sc_signal<bool> CAL0, CAL1, CAL2, CAL3, CAL4, CAL5 ;
sc_signal<bool> RET0, RET1, RET2, RET3, RET4, RET5 ;
mess2 ms2(“mess2”);
 ms2.cal2(CAL2);
ms2.ret2(RET2);
usecase1 uc1(“usecase1”);
uc1.arg4(ARG4);
 uc1.val5(VAL5);
 uc1.cal2(CAL2);
uc1.cal4(CAL4);
uc1.cal5(CAL5);
uc1.ret2(RET2);
uc1.ret4(RET4);
uc1.ret5(RET5);
usecase0 uc0(“usecase0”);
uc0.arg0(ARG0);
uc0.arg4(ARG4);
uc0.val0(VAL0);
uc0.val1(VAL1);
uc0.cal0(CAL0);
uc0.cal1(CAL1);
uc0.cal2(CAL2);
uc0.cal3(CAL3);
uc0.cal4(CAL4);
uc0.ret0(RET0);
uc0.ret1(RET1);
uc0.ret2(RET2);
uc0.ret3(RET3);
uc0.ret4(RET4);
return(0);}
```

4.3 Second technique

In this technique, we consider each end-to-end scenario as a VHDL process (SystemC SC_THREAD). An end-to-end scenario is a sequence of methods that are invoked by an external call from the environment. In this case, all processes communicate via shared variables. Table 2 shows the correspondence between UML and VHDL/SystemC concepts. All internal methods are implemented as VHDL procedures or functions. Since the same method may be called by many processes, we have to declare such methods globally in a VHDL package. We create ports only for external calls coming or returned values to the environment.

Table 2. Correspondence between UML and VHDL/SystemC for the second technique

UML concept	VHDL (behavioral /structural)	SystemC
End to end scenario	Process/Entity	SC_THREAD
Internal message without returned value	procedure	C++ function
Internal message with a returned value	function	C++ function
External call	port	port
Top level model	Test bench	sc_main()

By applying this technique on the above example, we obtain two VHDL processes: *process1* including the sequence of messages: *message_0*, *message_1*, and *message_2* and *process2* including *message_3*, *message_4*, *message_5*, and *message_2*. We observe that *message_2* is called by both *process_1* and *process_2*. Thus *message_2* is declared globally in a package. We use the *use* clause to import all messages defined in the package. *work* designates the user library where are stocked files resulting from VHDL code simulation.

```
package pack is
procedure message_2;
```

```
end package pack;

package body pack is
procedure message_2 is
begin
-- message_2 body
end procedure message_2;
end package body pack;
```

The VHDL behavioral style for the two processes is as follows:

```
entity  usecase_0 is
port (cal0, cal3 : inout boolean; arg0 : in integer; ret0, ret3 : out  boolean; val0 : out integer);
end entity usecase_0;

architecture system of usecase_0 is
library work;
use work.pack.all;
begin
process1 : process is
function message_1 return integer is
variable result : integer;
begin
-- message_1 body
message_2; -- call to message_2;
-- remaining instructions
return result;
end function message_1;

function message_0(arg : in integer) return integer is
variable ret1, result : integer;
begin
-- message_0 body
ret1 = message_1; -- call to message_1
return result;
end function message_0;
-- process code
variable arg;
begin
wait until cal0 = true;
cal0 <= false;
arg := arg0;
val0 <= message_0(arg)
ret0 <= true;
end process process1;

process2 : process is
function message_5 return integer is
variable result : integer;
begin
-- message_5 body
message_2; -- call to message_2;
-- remaining instructions
return result;
end function message_5;

procedure message_4 (arg : in integer) is
variable result : integer;
begin
-- message_4 body
Result := message_5; -- call to message_5;
-- remaining instructions
end procedure message_4;

procedure message_3 is
variable result arg : integer;
begin
-- message_3 body
arg := arg4;
result := message_4(); -- call to message_4;
-- remaining instructions
end procedure message_3;
begin
-- process code
begin
wait until cal3 = true;
cal3 <= false;
message_3;
ret3 <= true;
end process process2;
end architecture;
```

The VHDL structural style for the two processes is as follows:

```
entity proc1 is
port (cal0 : in boolean; arg0 : in integer; ret0 : out boolean; val0 : out integer);
end entity proc1;
architecture basic of proc1 is
begin
process1 : process is
-- process1 body
end process process1;
end architecture basis;

entity proc2 is
port (cal3 : in boolean; ret3 : out  boolean );
end entity proc2;
architecture basic of proc2 is
begin
process2 : process is
-- process2 body
end process process1;
end architecture basis;

architecture struct of usecase_0 is
begin
proces0 : entity work.proc0(basic)
    port map (cal0,arg0, ret0,val0);
proces1 : entity work.proc2(basic)
    port map (cal3,ret3);
end architecture struct;
```

The test bench architecture is the same as in the first technique. The corresponding SystemC code is as follows:

```
// system.h
# include "systemc.h"
SC_MODULE(system){
sc_in<int> arg0;
sc_inout<bool> cal0;
sc_inout<bool> cal3;
sc_out<bool> ret0;
sc_out<bool> ret3;
sc_out<bool> val0;
int message_0(int);
int message_1(void) ;
void message_2(void);
void message_3(void);
void message_4(int);
int message_5(void);
void process1();
void process2();
SC_CTOR(system) {
SC_THREAD(process1);
sensitive << cal0;
SC_THREAD(process2);
sensitive << cal3;
}};
// system.cc
void message_2(void){
// message_2 body}

int message_1(void){
// instructions
message_2() ;  // call to message_2
// remainig instructions}

int message_0(int) {
int result;
// instructions
Result = message_1();
// remaining instructions
return}

int message_5(void) {
// instructions
message_2() ;
// remaining instructions
Return}

void message_4(int) {
int result ;
// instructions
Result = message_5() ;
// remaining instructions}

void message_3(void) {
int arg ;
// instructions
if arg == 1 message_4(arg) ;
// remaining instructions}

void system::process1() {
wait();
cal0 = 0;
arg = arg0;
val0 = message_0(arg);
ret0 = 1; }
void system::process2() {
wait();
cal3 = 0;
message_3();
ret3 = 1; }

// main.cc
#include "system.h"
int  sc_main(int argc, char* argv[]) {
sc_signal<bool> CAL0, CAL3, RET0, RET3;
sc_signal<int> ARG0,VAL0;
system  sys("system");
sys.arg0(ARG0);
sys.cal0(CAL0);
sys.cal3(CAL3);
sys.ret0(RET0);
sys.ret3(RET3);
sys.val0(VAL0);
return(0); }
```

4.4 Third technique

In this technique, each UML object is considered as a VHDL (SC_THREAD) process. For each input /output message call, we create input/output ports (we add more ports for arguments and returned values). Table 3 shows the correspondence between UML and VHDL/SystemC concepts.

Table 3. Correspondence between UML and VHDL/SystemC for the third technique

UML concept	VHDL (behavioral /structural)	SystemC
Object	Process/Entity	SC_THREAD
Input message call	Input ports	Input ports
Output message call	Output ports	Output ports
External call	port	port
Top level model	Test bench	sc_main()

By applying this technique on the above example, we obtain four processes (4): *Env*, *class_0*, *class_1*, and *class_2*. For the sake of the space, we give only the VHDL code for *Env* and *class_0*.

```
entity  usecase_0 is
port (cal0, cal1, cal2, cal3, cal4, cal5 : inout boolean; arg0, arg4 : inout integer; ret0, ret1, ret2, ret3, ret4, ret5 : inout boolean; val0, val1, val5 : inout integer);
end entity usecase_0;
```

```
architecture system of usecase_0 is
begin
Env : process is
variable temp : integer;
begin
cal0 <= true;
arg0 <= 1;
wait until ret0 = true;
ret0 <= false;
temp := val0;
--code
cal3 <= true;
wait until ret3 = true;
ret3 <= false;
-- remaining code
end process Env;

class_0 : process is
variable  arg, temp : integer;
begin
wait until cal0 = true;
cal0 <= false;
arg := arg0;
-- message0 instructions
cal1 <= true;
wait until ret1 = true;
ret1 <= false;
-- remaining message_0 instructions
ret0 <= true;
val0 <= w;
wait until cal3 = true;
cal3 <= false;
-- message3 instructions
temp := a;
if temp = 1 then
cal4 <= true;
wait until ret4 = true;
ret4 <= false;
end if
-- remaining message_3 instructions
ret3 <= true;
end process class_0;
end architecture system;
```

For the sake of space, we show only the structure of the *Env* process:

```
entity  Environment is
port (cal0, cal3 : out boolean; ret0, ret3 : inout  boolean;
arg0 : out integer; val0 : in integer);
end entity Environment;

architecture basic of Environment is
begin
Env : process is
-- Env process code
end process Env;
end architecture basic;
```

```
architecture struct of usecase_0 is
signal cal0, cal1, cal2, cal3, cal4, cal5 : boolean;
signal arg0, arg4, val0, val1, val5 : integer;
begin
Envr : entity work.Environment(basic)
       port map (cal0, cal3, ret0, ret3, arg0, val0);
clas0 : entity work.class0(basic)
       port map (cal0, cal1, cal3, cal4, ret0, ret1, ret3, ret4,
arg0, arg4, val0, val1);
clas1 : entity work.class1(basic)
       port map (cal1, cal2, cal4, cal5, ret1, ret2, ret4, ret5,
arg4, val1, val5);
clas2 : entity work.class2(basic)
       port map (cal2, cal5, ret2, ret5, val5);
end architecture struct;
```

For the sake of space, we give only the SystemC code for *Env* and *class_0*.

```
// system.h
# include "systemc.h"
SC_MODULE(system){
sc_inout<bool> cal0 ;
sc_inout<bool> cal1;
sc_inout<bool> cal2;
sc_inout<bool> cal3;
sc_inout<bool> cal4;
sc_inout<bool> cal5;
sc_inout<bool> ret0;
sc_inout<bool> ret1;
sc_inout<bool> ret2;
sc_inout<bool> ret3;
sc_inout<bool> ret4;
sc_inout<bool> ret5;
sc_inout<int> arg0, arg4,val0, val1, val5;
void env();
void class_0();
void class_1();
void class_2();
SC_CTOR(system) {
SC_THREAD(env);
sensitive << ret0 << ret3 ;
SC_THREAD(class_0);
sensitive << cal0 << ret1 << cal3 << ret4 ;
SC_THREAD(class_1);
sensitive << cal1 << ret2 << cal4 << ret5 ;
SC_THREAD(class_2);
sensitive << cal5 << cal2 ;}};
// system.cc
#include "system.h"
void system::env() {
int temp;
cal0 = 1;
arg0 = 1; // some initialization
wait (ret0);
ret0 = 0;
temp = val0;
cal3 = 1;
wait (ret3);
ret3 = 0;
}
void system::class_0() {
int arg, temp;
wait (cal0);
cal0 = 0;
arg = arg0;
-- message0 instructions
cal1 = 1;
wait (ret1);
ret1 = 0;
-- remaining message_0 instructions
ret0 = 1;
Val0 = w;
wait (cal3);
cal3 = 0;
-- message3 instructions
temp := a;
if temp = 1{
cal4 = 1;
wait (ret4);
ret4 = 0;}
-- remaining message_3 instructions
ret3 = 1;
}
void system::class_1() {
// body of class_1
}
void system::class_2() {
// body of class_2
}
// main.cc
#include "system.h"
int  sc_main(int argc, char* argv[]) {
sc_signal<bool> CAL0, CAL1, CAL2, CAL3, CAL4, CAL5;
sc_signal<bool> RET0, RET1, RET2, RET3, RET4, RET5;
sc_signal<int> ARG0,ARG4,VAL0,VAL1, VAL5;
system sys("system");
sys.arg0(ARG0);
 sys.arg4(ARG4);
 sys.val0(VAL0);
sys.val1(VAL1);
sys.val5(VAL5);
sys.cal0(CAL0);
sys.cal1(CAL1);
sys.cal2(CAL2);
sys.cal3(CAL3);
sys.cal4(CAL4);
sys.cal5(CAL5);
sys.ret0(RET0);
sys.ret1(RET1);
sys.ret2(RET2);
sys.ret3(RET3);
sys.ret4(RET4);
sys.ret5(RET5);
return(0) ;}
```

Table 4 compares between the three techniques.

Table 4. Comparison between the three techniques

Technique	Processes Number	Process Granularity	Communication scheme
First	6	Fine	Message Passing
Second	2	Coarse	Shared memory
Third	4	Coarse	Mix

4.5 Modeling with UML activity diagrams

In our proposed flow (see figure 1), the second step consists in internal behaviour modelling of messages using UML activity diagrams whose state actions are expressed in the Action Language (AL) included in the Rhapsody environment. The AL is a subset of C++ that uses a C++ compiler to enable the model simulation. This language provides message passing, data checking, actions on transitions, and model execution. It supports majority of C++ operators, *if/else*, *for*, *while*, *do/while*, *return* instructions, primitive types, array of primitives, objects, invoking block operations, generating events, generating port events, testing port for an event, etc…figure 6 shows an example of an UML activity diagram with an action including three assignments written in AL, a call to a message called *Message_1* belonging to *class_0*, a condition, and a termination state. Note that in our case, only a subset of the AL is used. For instance, pointers are not used since existing Hardware synthesize tools do not know synthesize pointers to hardware. Instead of, we use arrays. VHDL supports a large set of operators and control instructions found in AL. Using the Rhapsody environment we can perform functional simulation before HDL code generation. This step is very important in order to validate the HDL code functionality against UML functional models.

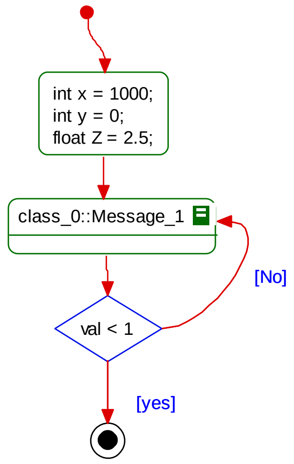


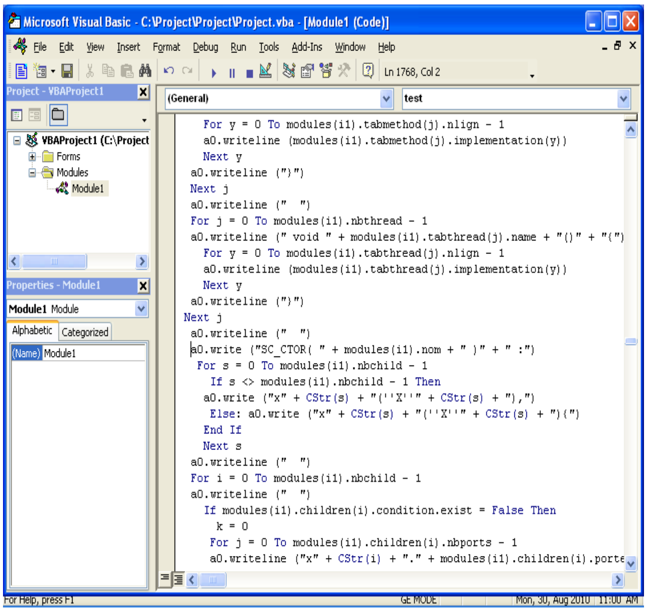
Figure 6. Example of UML activity diagram

5. Implementation and case study

We have used the Rhapsody environment for UML modelling and HDL code generation. In order to automate the VHDL/SystemC code generation from UML models, we have used the VB API which is integrated in the Rhapsody environment. With VB, we can easily parse UML graphical models then collect the necessary information to create VHDL/SystemC files (see figure 7). We have developed two VB macros for SystemC/VHDL codes generation and integrated them as tool boxes in the Rhapsody environment. As a case study, we have chosen the SDP (Simplex Data Protocol) [19] application whose UML sequence diagrams are illustrated in figure 8. Figure 9 shows the UML activity diagram for the receiver object. Figure 10 gives us an overview of SystemC files for the *receiver* object.

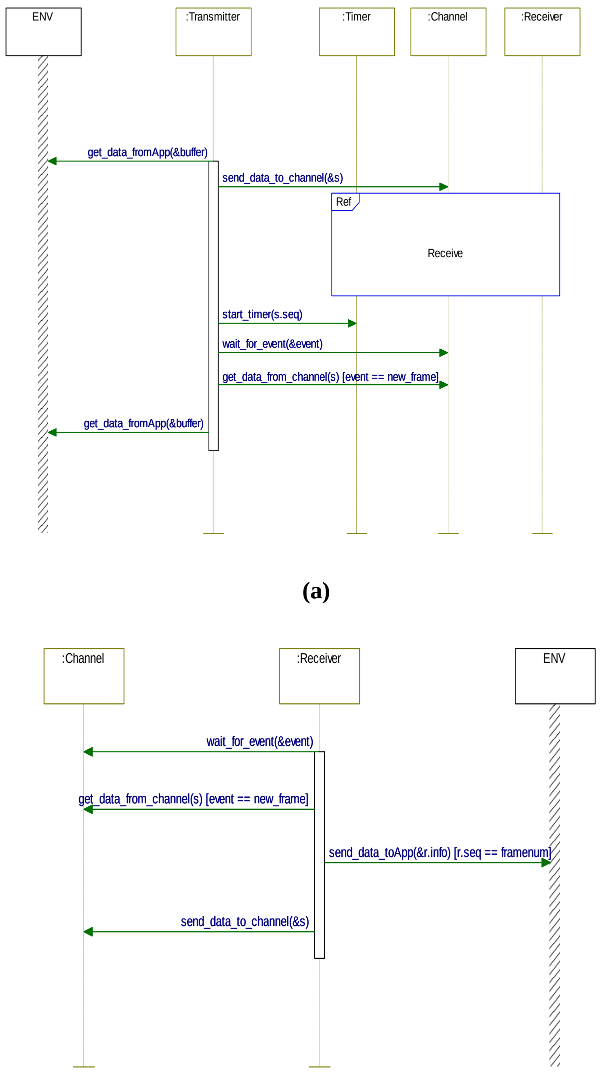
6. Conclusion

In this paper, we present our approach for automatic VHDL/SystemC code generation from UML models at early stages of embedded systems development. Our proposed flow consists mainly of two steps: generation of VHDL/SystemC codes from UML hierarchic sequence diagrams then from UML activity diagrams. The generated VHDL/SystemC code at the first stage is used for algorithmic space exploration and simulation purposes using existing commercial simulators. In the second step, we introduce UML activity diagrams to model messages internal behaviours. Actions of activity diagrams are expressed in the C++ Action Language (AL) which is included in the Rhapsody environment. From AL, a full VHDL/SystemC code is generated for both simulation and synthesis. VHDL/SystemC code is generated as text files automatically and this is due to the VB API included in the Rhapsody environment. In order to enable designer to explore the algorithmic space, we proposed three techniques for HDL code generation. According to simulation results, the designer can restructure his/her system by increasing or decreasing the processes number (i.e. merge or scatter processes). As a perspective, we plan to investigate the MDA approach for VHDL/SystemC code generation from sequence diagrams and consider asynchronous events and temporal constraints.



```
    For y = 0 To modules(i1).tabmethod(j).nlign - 1
    a0.writeline (modules(i1).tabmethod(j).implementation(y))
    Next y
  a0.writeline ("}")
  Next j
  a0.writeline ("  ")
  For j = 0 To modules(i1).nbthread - 1
  a0.writeline (" void " + modules(i1).tabthread(j).name + "()" + "{")
    For y = 0 To modules(i1).tabthread(j).nlign - 1
    a0.writeline (modules(i1).tabthread(j).implementation(y))
    Next y
  a0.writeline ("}")
 Next j
  a0.writeline ("  ")
  a0.write ("SC_CTOR( " + modules(i1).nom + " )" + " :")
   For s = 0 To modules(i1).nbchild - 1
     If s <> modules(i1).nbchild - 1 Then
    a0.write ("x" + CStr(s) + "(''X''" + CStr(s) + "),")
     Else: a0.write ("x" + CStr(s) + "(''X''" + CStr(s) + "){")
    End If
    Next s
  a0.writeline ("  ")
  For i = 0 To modules(i1).nbchild - 1
  a0.writeline ("  ")
    If modules(i1).children(i).condition.exist = False Then
      k = 0
     For j = 0 To modules(i1).children(i).nbports - 1
     a0.writeline ("x" + CStr(i) + "." + modules(i1).children(i).porte
```

Figure 7. Programming with VB API


ENV
:Transmitter
:Timer
:Channel
:Receiver
get_data_fromApp(&buffer)
send_data_to_channel(&s)
Ref
Receive
start_timer(s.seq)
wait_for_event(&event)
get_data_from_channel(s) [event == new_frame]
get_data_fromApp(&buffer)
:Channel
:Receiver
ENV
wait_for_event(&event)
get_data_from_channel(s) [event == new_frame]
send_data_toApp(&r.info) [r.seq == framenum]
send_data_to_channel(&s)

(a)

(b)

Figure 8. UML sequence diagrams for SDP

(a) Main sequence diagram; (b) sequence diagram for receive use case.

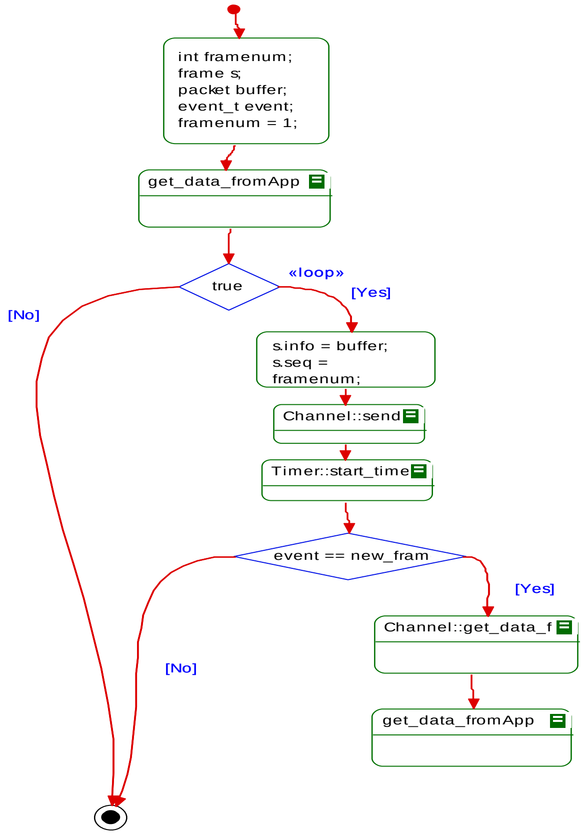
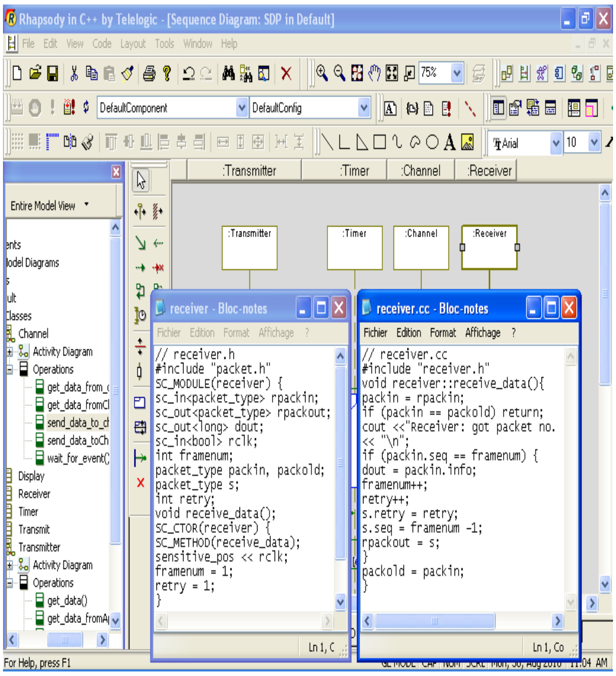


Figure 9. UML activity diagram for Receiver object



```
// receiver.h
#include "packet.h"
SC_MODULE(receiver) {
sc_in<packet_type> rpackin;
sc_out<packet_type> rpackout;
sc_out<long> dout;
sc_in<bool> rclk;
int framenum;
packet_type packin, packold;
packet_type s;
int retry;
void receive_data();
SC_CTOR(receiver) {
SC_METHOD(receive_data);
sensitive_pos << rclk;
framenum = 1;
retry = 1;
}
```

```
// receiver.cc
#include "receiver.h"
void receiver::receive_data(){
packin = rpackin;
if (packin == packold) return;
cout <<"Receiver: got packet no.
<< "\n";
if (packin.seq == framenum) {
dout = packin.info;
framenum++;
retry++;
s.retry = retry;
s.seq = framenum -1;
rpackout = s;
}
packold = packin;
}
```

Figure 10. SystemC code generation from Rhapsody UML models

References

[1] Akehurst, D.H., Uzenkov, O., Howells, W.G., Mcdonald Maier, K.D., Bordbar, B., "Compiling UML state diagrams into VHDL: An experiment in Using Model Driven Development", FDL'07, 2007.

[2] Ashenden, P.J., "The VHDL cookbook", first edition, published by Morgan Kaufmann, 1990.

[3] Ashenden, P.J., "VHDL Tutorial", Elsevier Science (USA), 2004.

[4] Bjarklund, D., Lilius, J., Poress, I., "Towards efficient code synthesis from Statecharts", Puml Workshop at UML2001, 2001.

[5] Bjarklund, D., Lilius, J., "From UML behavioral descriptions to efficient synthesizable VHDL", proceedings of 20th IEEE NORCHIP Conference, Copenhagen, Denmark, 2002.

[6] Booch, G., Rumbaugh, J., Jacobson I., "Unified Modeling Language User Guide", Addison-Wesley, 1999.

[7] Boutekkouk, F., Benmohammed, M., Bilavarn, S., Auguin, M., "UML2.0 profiles for Embedded Systems and Systems On a Chip (SOCs)", JOT (Journal of Object Technology), January, 2009.

[8] Coyle, F.P, Thornton, M.A., "From UML to HDL: a Model Driven Architectural Approach to Hardware-Software Co-Design", proceedings of Information Systems: New Generations Conference (ISNG), p. 88-93, 2005.

[9] Gajski, D., Vahid, F., Narayan, S., Gong, J., "Specification and Design of Embedded Systems", Prentice Hall. Englewood, New jersey 07632, 1994.

[10] Jerraya, A.A., Wolf, W., "Multiprocessor systems on chip", Morgan Kaufmann publishers, 2005.

[11] Martin G., "UML for embedded systems specification and design: motivation and overview", Design, Automation and Test in Europe Conference and Exhibition, 2002. Proceedings, p. 773–775, 2002.

[12] McUmber, W.E., Cheng, B.H.C., "UML-based analysis of embedded systems using a mapping to VHDL", proceedings of IEEE Int. Symposium on High Assurance Software Engineering (HASE'99), Washington, DC, USA, p. 56-63, 1999.

[13] ModelSim documentation, ftp://ftp.xilinx.com/pub/documentation.

[14] Narayan, S., Vahid, F., Gajski, D.D, "Translating system specifications to VHDL", IEEE European Design Automation Conference, Amsterdam, Netherlands, 1991.

[15] Rhapsody UML modeler from Telelogic, an IBM company. www.telelogic.com/products/rhapsody

[16] Riccobene, E., Scandura, P., Rosti, A., Bocchino, S., "A SOC Design Methodology Involving a UML2.0 Profile for SystemC", Proceedings of the Design, Automation and Test in Europe Conference end Exhibition (DATE'05), 2005.

[17] Rieder, M., Steiner, R., Berhouzoz, C., Corthay, F., Sterren, T., "Synthesized UML, a Practical Approach to Map UML to VHDL", LNCS, Volume 3943, 2006.

[18] SystemC, Functional specification for SystemC 2.0, www.systemc.org, 2002.

[19] SystemC, Version 2.0 User's guide, www.systemc.org, 2002.

[20] SystemC, IEEE Standard SystemC® language Reference Manual, www.systemc.org, 2005.

[21] Thomson, R., Chouliaras, V., Mulvaney, D., Plessis, P., "From UML to Structural Hardware Designs", UMLSOC, 2007.

[22] UML2.0 Superstructure Specification, http://www.omg.org, 2003.

[23] VHDL, IEEE Standard VHDL Language Reference Manual. IEEE, IEEE Std 1076, 2000.

[24] Vidal, J., De Lamotte, F., Gogniat, G., Soulard, P., Diguet, JP., "A codesign approach for embedded system modeling and code generation with UML and MARTE", DATE09, 2009.

Author Biography

Fateh Boutekkouk was born in Constantine (Algeria). He received his BS degree in Computer science from the University of Constantine, his MS degree from the University of Jijel (Algeria), and his PhD from the University of Constantine in 2010. He is an assistant professor at the University of Oum el Bouaghi (Algeria) since 2003. His current research interests include Software Engineering, Embedded Systems and Systems On Chip (SOC) design.

An Energy-Efficient Unicast Routing Protocol for Wireless Sensor Networks

Young-Jun Chung[1]

[1]Department of Computer Science, Kangwon National University, Chunchon, Korea
e-mail: ychung@kangwon.ac.kr

***Abstract*:** The efficient node-energy utilization in wireless sensor networks is very important because sensor nodes operate with limited battery power. To increase the lifetime of the wireless sensor networks, we reduced the node energy consumption of the overall network while maintaining all sensors balanced node power use. Since a large number of sensor nodes are densely deployed and interoperated in wireless sensor network, the lifetime extension of a sensor network is maintained by keeping many sensor nodes alive. In this paper, we present an energy-efficient unicast routing protocol for wireless sensor networks to increase its lifetime without degrading network performance. The proposed protocol is designed to avoid traffic congestion on specific nodes at data transfer and to make the node power consumption widely distributed to increase the lifetime of the network. The performance of the proposed protocol has been examined and evaluated with the NS-2 simulator in terms of network lifetime and end-to-end delay.

***Keywords*:** wireless sensor network, energy-efficient unicast routing protocol, NS-2

1. Introduction

A wireless sensor network is one of the ad hoc wireless telecommunication networks, which are deployed in a wide area with tiny low-powered smart sensor nodes. An essential element in this ubiquitous environment, this wireless sensor network can be utilized in a various information and telecommunication applications. The sensor nodes are small smart devices with wireless communication capability, which collects information from light, sound, temperature, motion, etc., processes the sensed information and transfers it to other nodes.

A wireless sensor network is typically made of many sensor nodes for sensing accuracy and scalability of sensing areas. In such a large scale of networking environment, one of the most important networking factors are self-organizing capability for well adaptation of dynamic situation changes and interoperating capability between sensor nodes[1]. Many studies have shown that there are a variety of sensors used for gathering sensing information and efficiently transferring the information to the sink nodes.

The major issues of such studies are protocol design in regards to battery energy efficiency, localization scheme, synchronization, data aggregation and security technologies for wireless sensor networks. In particular, many researchers have great interest in the routing protocols in the network layer, which considers self-organization capabilities, limited batter power, and data aggregation schemes[2][3].

A wireless sensor network is densely deployed with a large number of sensor nodes, each of which operates with limited battery power, while working with the self-organizing capability in the multi-hop environment. Since each node in the network plays both terminal node and routing node roles, a node cannot participate in the network if its battery power runs out. The increase of such dead nodes generates many network partitions and consequently, normal communication will be impossible as a sensor network. Thus, an important research issue is the development of an efficient batter-power management to increase the life cycle of the wireless sensor network [4].

In this paper, we proposed an efficient energy aware routing protocol, which is based upon the on-demand ad hoc routing protocol AODV[5][6], which determines a proper path with consideration of node residual battery powers. The proposed protocol aims to extend the life time of the overall sensor network by avoiding the unbalanced exhaustion of node batter powers as traffic congestion occurs on specific nodes participating in data transfer.

In section 2 of this paper, we describe the well-known AODV routing protocol and show some difficulties in adapting the protocol for wireless sensor network. In section 3, we propose an efficient routing protocol, which considers the node residual battery power while extending the life time of the network. Section 4 discusses the NS-2 simulation performance analysis of the routing protocols along with final conclusions and future studies.

2. Related Studies

The AODV(Ad hoc On-demand Distance Vector) protocol is an on-demand routing protocol, which accomplishes the route discovery whenever a data transfer is requested between nodes. The AODV routing protocol searches a new route only by request of source nodes. When a node requests a route to a destination node, it initiates a route discovery process among network nodes. The protocol can greatly reduce the number of broadcasts requested for routing search processes, when compared to the DSDV (Destination Sequenced Distance Vectors) routing protocol, which is known to discover the optimum route between source and destination with path information of all nodes. Additionally, since each node in the DSDV routing protocol maintains a routing table - data which includes complete route information - the AODV protocol greatly improves some drawbacks of DSR (Dynamic Source Routing) protocol such as the overhead incurred at data transfer.

Once a route is discovered in the AODV routing protocol, the route will be maintained in a table until the route is no longer used. Each node in the AODV protocol contains a sequence number, which increases by one when the location of a neighbor node changes. The number can be used to determine the recent route at the routing discovery.

The AODV protocol utilizes a similar routing discovery process as the DSV protocol but uses a different process to maintain and manage a routing table. The nodes of the DSV protocol maintains all routing information between source and destination but the nodes of the AODV protocol have path information in a brief routing table, which stores the destination address, destination sequence number, and next hop address.

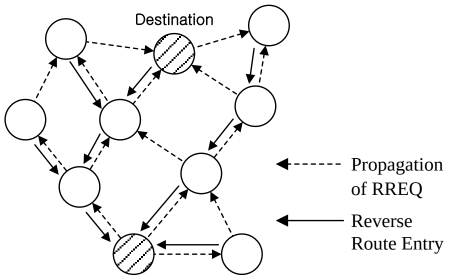


Figure 1. Flooding of RREQ messages

Each entry of a routing table has a lifetime field which is set when its routing information is updated and changed. An entry will be removed from the routing table when its lifetime is expired. Moreover, to maintain a routing table, the AODV protocol periodically exchanges routing messages between neighbor nodes. Such processes typically raise significant overhead and wastes available bandwidth. However, the AODV protocol reduces the latency time of the routing discovery and determines efficient routes between nodes.

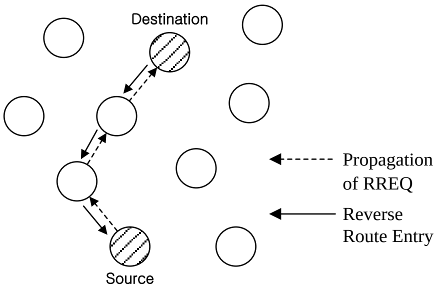


Figure 2. A routing establishing flow between source and destination

The route discovery process of the AODV protocol is similar to that of DSR. A source node broadcasts a RREQ (Route REQquest) packet to find a route to a destination node. When a neighbor node receives the RREQ packet, it rebroadcasts the packet to intermediate nodes until the packet arrives at a destination node. At the same time, the intermediate node or the destination node, which receives a RREQ packet, replies a RREP (Route reply) packet back to the source node. The destination node collects all RREQ messages during a time interval, determines a least hop-count route, and then sends a RREP message to the source node.

The sequence number of a RREQ packet can eliminate a loop generation and make an intermediate nodes reply only on recent route information. When an intermediate node forwards a RREQ packet to neighbor nodes, the receiver node records the intermediate node into the routing information in order to determine the forwarding path. Such processes repeat until arriving at the destination. Then the destination node sends a RREP message, which includes the routing, to the source via the reverse path. In the case that a node receives duplicated RREQ messages, it uses only the first message and ignores the rest. If errors occur on a specific link of the routing path, either a local route recovery process is initiated on a related node or a RERR(Route Error) message will be issued to the source for a source route recovery process. In such cases, the intermediate nodes receiving the RERR message eliminates all routing information related to the error link.

The AODV routing protocol determines a least hop-count path between a source and a destination, thus minimizing the end-to-end delay of data transfer. Since the protocol uses the shortest route for end-to-end data delivery, it minimizes the total energy consumption.

However, if two nodes perform data transfer for long time on the specific path, nodes belonging in this path use more battery power than other nodes, resulting in earlier powering out of nodes. The increase of power-exhausted nodes creates partitions in the wireless sensor network. The nodes belonging to these partitions cannot transfer any further data, thus killing the lifetime of the network.

In order to extend the lifetime of the network, one possible solution is to make equally balanced power consumption of sensor nodes. Since AODV routing mechanism does not consider the residual energy of nodes at the routing setup, and since it considers only routing hop count as a distance metric, such unbalanced node energy consumptions occurs. An efficient routing algorithm is proposed, which considers both node hop-count and node energy consumption in section 3.

3. Proposed Routing Protocol

In this paper, we describe a routing protocol, which considers a residual energy of sensor nodes to avoid unbalanced energy consumption of sensor nodes. The proposed protocol is based upon a reactive ad hoc AODV routing algorithm. The protocol can make the node energy consumption balanced and extend overall network lifetime without performance degradation such as delay time, compared to the AODV routing algorithm.

3.1 Operations of the proposed routing protocol

The proposed protocol performs a route discovery process similar to the AODV protocol. The difference is to determine an optimum route by considering the network lifetime and performance; that is, considering residual energy of nodes on the path and hop count. In order to implement such functions, two new fields, called a Min-RE(Minimum Residual Energy) field and a TRE(Total Residual Energy) field , are added to the RREQ message as shown in Figure 3. The Min-RE field and the TRE field are set as a default value of -1 and 0, respectively, when a source node broadcasts a new RREQ message for a route discovery process.

To find a route to a destination node, a source node floods a RREQ packet to the network. When neighbor nodes receive the RREQ packet, they update the Min-RE value and the TRE value and rebroadcast the packet to the next nodes until the packet arrives at a destination node.

Type	J	R	G	D	U	Reserved	Hop Count
RREQ ID.							
Destination IP Address							
Destination Sequence Number							
Originator IP Address							
Originator Sequence Number							
Min-RE(Added)							
TRE(Added)							

Figure 3. A RREQ message format for our proposed protocol

If the intermediate node receives a RREQ message, it increases the hop count by one and replaces the value of the Min-RE field with the minimum energy value of the route. In other words, Min-RE is the energy value of the node if Min-RE is greater than its own energy value; otherwise Min-RE is unchanged. The update algorithm is shown in Figure 4.

```
  Update_RREQ ()
  {
    If Node receives RREQ message then
    {
       RREQ.Hop_count ← RREQ.Hop_count +1;
       RREQ.TRE ← RREQ.TRE + Res_Energy;
       If (RREQ.Min.RE=1 or
RREQ.Min_RE>Res_Energy)
       then RREQ.Min_RE ← Res_Energy;
     }
     Rebroadcast RREQ;
   }
```

Figure 4. A RREQ update algorithm at an intermediate node

```
  Reply_RREP()
  {
   If first RREQ message received the
   {
     Start timer_RREQ;
     i ← 0;
     while (timer_RREQ is not expired) do
     {
    //compute ARE,α and store forwarding node ID for each
route
    ARE ← RREQ.TRE/(RREQ.Hop_count -1;
     α[i]←
(RREQ.Min_RE*p+ARE*(1-p))/(RREQ.Hop_count);
    forwarding_node[i] ← RREQ.forwarding_node _ID;
    i ← i +1;
   wait for next RREQ;
  }
  Find index value n for maximum α value;
  Send RREP message to forwarding_node[n];
 }
}
```

Figure 5. A path selection algorithm at the destination node

Although intermediate nodes have route information to the destination node, they keep forwarding the RREQ message to the destination because it has no information about residual energy of the other nodes on the route. If the destination node finally receives the first RREQ message, it triggers the data collection timer and receives all RREQ messages forwarded through other routes until time expires. After the destination node completes route information collection, it determines an optimum route with use of a formula shown in 3.2 and then sends a RREP message to the source node by unicasting. Figure 5 shows the path selection algorithm. If the source node receives the RREP message, a route is established and data transfer gets started. Such route processes are performed periodically, though node topology does not change to maintain node energy consumption balanced. That is, the periodic route discovery will exclude the nodes having low residual energy from the routing path and greatly reduce network partition.

3.2 Determination of routing

The optimum route is determined by using the value of α described in formula (1). The destination node calculates the values of α for received all route information and choose a route that has the largest value of α. That is, the proposed protocol collects routes that have the minimum residual energy of nodes relatively large and have the least hop-count, and then determines a proper route among them, which consumes the minimum network energy compared to any other routes.

$$\alpha = \frac{Min\_RE \times p + ARE \times (1-p)}{\#Hops} \quad \text{...........} \ (1)$$

Here Min-RE is the minimum residual energy on the route and No-Hops is the hop count of the route between source and destination. ARE is the average node residual energy on the path. And p is a weight coefficient to adjust the ration of Min_RE and ARE. When p goes to 1, most nodes having less energy are removed from the optimum path selection. This case is used when an energy difference between nodes is large. Inversely, when p goes to 0 the optimum path selection will be determined by only the hope count and ARE

3.3 The analysis of routing protocols

To understand the operations of the proposed protocol, we consider three different routing protocols for operational comparison:

- Case 1: Choose a route with the minimum hop count between source and destination.
(AODV routing protocol).
- Case 2: Choose a route with largest minimum residual energy. (Max_Min Energy (Min-ER) routing protocol)
- Case 3: Choose a route with the large minimum residual energy and less hop count. i.e. with the longest network lifetime (our proposed routing protocol).

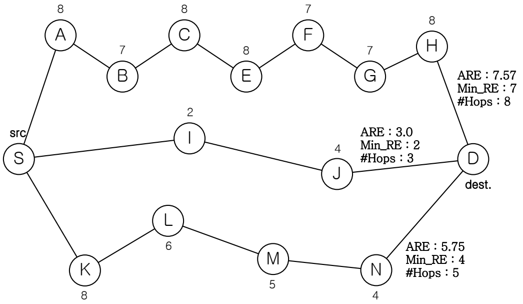


Figure 6. A sample network for establishment of routing paths

Consider a network illustrated in Figure 6. Here we consider a simple routing example to setup route from source node S to destination node D. The number written on a node represents the value of residual node energy. We consider three different cases of routes. Since the Case 1 considers only the minimum hop count, it selects route <S-I-J-D> which has the hop count of 3. In the Case-2, select route <S-A-B-C-E-F-G-H-D> which has Min-RE 7 is chosen because the route has the largest minimum residual energy among routes. Our proposed model needs to compute the value of α by using formula (1), and selects a route with largest value of α. Thus Case 3 selects route <S-K-L-M-N-D> which has largest α value of 0.975 with p=.5.

Case 1 selects the shortest path without considering residual energy of nodes, which is the same as the AODV routing algorithm. This case does not sustain a long lifetime in the network as described in section II. Case 2 selects a route with largest minimum residual energy to extend network lifetime but it has serious problem in terms of the hop count. Case-3 improves the drawbacks of Case 1 and Case-2 by considering both residual energy and hop count. It extends network lifetime by arranging almost all nodes to involve in data transfer. The proposed protocol also selects a route with the longest lifetime in the network without performance degradation such as delay time and node energy consumption.

4. Performance Evaluation

The performance analysis of routing protocols is evaluated with the NS-2 simulator[7]. Then our proposed protocol is compared to other two routing protocol (Case 1 and Case 2) in terms of the average end-to-end delay and the network lifetime.

4.1. Simulation Environment

In this simulation, our experiment model performed on 100 nodes which were randomly deployed and distributed in a 500×500 square meter area. We assume that all nodes have no mobility since the nodes are fixed in applications of most wireless sensor networks. Simulations are performed for 60 seconds. We set the propagation model of wireless sensor network as two-ray ground reflection model and set the maximum transmission range of nodes as 100 meters. The MAC protocol is set to IEEE 802.11 and the bandwidth of channel is set to 1Mbps.

Each sensor node in the experimental network is assumed to have an initial energy level of 7 Joules. A node consumes the energy power of 600mW on packet transmission and consumes the energy power of 300mW on packet reception. The used traffic model is an UDP/CBR traffic model. Size of data packet is set to 512byte and traffic rate varies to 2, 3, 4, 5, 6, 7, 8, 9, 10 packets/sec to compare performance depend on traffic load. In this simulation, the weight coefficient k is calculated based on traffic model, bandwidth, and energy consumption of a node. Our simulation model uses a sensor network that has the bandwidth of 1 Mbps, the packet size of 512 bytes. Thus, packet transmission time per link is calculated as about 0.004096seconds and the node energy consumption for our simulation model is about 0.0037 Joule.

4.2. Simulation Results

The major performance metrics of a wireless sensor network are the end-to-end delays (or throughput) and network lifetime. In order to compare network lifetime of three different routing protocols, we measured the number of exhausted energy nodes every second for 60 seconds. Figure 5 illustrates that number of exhausted node of each model according to simulation time. The vertical axis is represented the number of exhausted energy nodes in the network. The increase of the exhausted energy nodes may cause a network partition that makes network functions impossible. The number of exhausted energy nodes in AODV (Case 1), Min-ER (Case 2), and our protocol start appearing at 35, 42, and 47 seconds, respectively. The number in these protocols is saturated on 80% of nodes at 45, 48, and 55 seconds, respectively. As shown in Figure 7, our proposed protocol has longer lifetime duration than other protocols. In Particular, 60% of nodes in our protocol work normally at the elapsed time of 55 seconds compared to 20 % in other protocols. This result shows that our routing protocol properly leads to balanced energy consumption of sensor nodes.

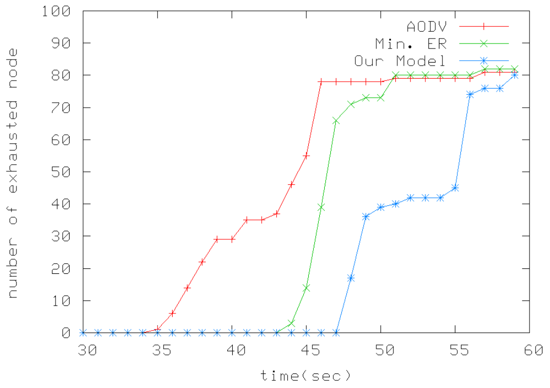


Figure 7. Comparison of the number of exhausted energy nodes

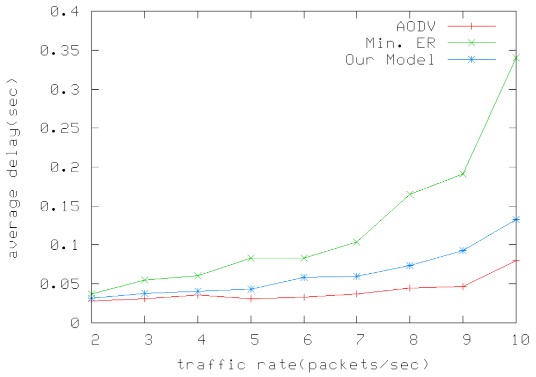


Figure 8. End-to end delay for traffic rate

Figure 8 gives the average end-to-end delay of all three protocols in respect with traffic loads. The AODV protocol has minimum delay and Min-ER has maximum delay. Additionally, the delay of our protocol was little higher than that of AODV. Our protocol has a relatively good delay characteristic without degradation of performance compared to AODV.

Based upon the simulation results, we confirmed that our proposed protocol can control the residual node energy and the hop count in a wireless sensor network and effectively extend the network lifetime without performance degradation.

5. Conclusions

In this work, we proposed an energy efficient unicast routing protocol which improves the lifetime of sensor networks. The protocol considers both hop count and the residual energy of nodes in the network. Based upon the NS-2 simulation, the protocol has been verified with very good performance in network lifetime and end-to-end delay. If we used a simulation mode of the large number of nodes (or 1000 or more), our protocol make network lifetime much longer compared to AODV and Min-ER protocols. Consequently, our proposed protocol can effectively extend the network lifetime without other performance degradation.

The applications in wireless sensor networks may require different performance metrics. Some applications are focused on the lifetime of network and the others on delay. Some efficient routing mechanisms in respect with applications may be needed for further studies.

References

[1] Ian F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” IEEE Communications Magazine, volume 40, Issue 8, pp.102-114, Aug. 2002.

[2] K. Akkaya and M. Younis, "A Survey of Routing Protocols in Wireless Sensor Networks, " in the Elsevier Ad Hoc Network Journal, Vol 3/3, pp.325-349, 2005.

[3] Q. Jiang and D. Manivannan, “Routing protocols for sensor networks,” Proceedings of CCNC 2004, pp.93-98, Jan. 2004.

[4] Suresh Singh and Mike Woo, "Power-aware routing in mobile ad hoc networks", Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking, Dallas, Texas, pp. 181 -190, 1998.

[5] Charles E. Perkins and Elizabeth M. Royer. "Ad hoc On-demand Distance Vector Routing." Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, pp. 90-100, February 1999.

[6] Charles E. Perkins, "Ad hoc On-demand Distance Vector (AODV) Routing.", RFC 3561, IETF MANET Working Group, July 2003.

[7] Information Sciences Institute, "The Network Simulator ns-2" http://www.isi.edu/nanam/ns/, University of Southern California.

Author Biography

Young J. Chung received the BS degree in Electrical Engineering from Seoul National University, Korea, in 1974 and the MS and Ph.D. degrees in electrical and computer engineering from the University of Kansas, Lawrence, Kansas, in 1983 and 1988, respectively. Between May 1982 and May 1988, he worked with Advanced Computer Division at KGS, Lawrence, Kansas, USA. Between May 1988 and August 1991, he worked with Computer Division at RANPAC, Rancho California, California, USA. Since August 1991 he has been a faculty at the Department of Computer Science at Kangwon National University, Korea. His research interests are in the areas of wireless and mobile communications systems, wireless sensor networks, Internet applications and services, network security.

The Impact of the Digital Divide on Education and Health

Veena Paliwal[1]and Vaibhav Sharma[2]

[1]University of Illinois at Urbana-Champaign, 61820 IL, USA
[2]Winthrop University, Rock Hill, 29733 SC, USA
paliwal2@illinois.edu, sharmavai@winthrop.edu

Abstract: *Digital divide, the gap between technology haves and haves-not, is rising in all areas where the use of computers is prominent. This gap gives rise to several related problems like poverty, unemployment, illiteracy, and fewer health benefits. This article examines the existence of digital divide and discusses its impact on life, education, and health. Digital divide in education is an acute problem as those with awareness of information and computing technologies (ICT) are reaping the benefits of online education and all available resources that not only improve productivity but makes information easily accessible. Digital divide in the health education is rising and people with access to ICT are getting all useful health information from the internet and improving the quality of life. The issue of health gives rise to double divide since economic and social disparities reinforced divide on those who can't afford to have computers or internet access.*

Keywords**:** *digital divide, health, information and computing technologies*.

1. Introduction

The term digital divide (DD) refers to the gap between those who have access to information and technology and those who do not. Since the use of computers is wide-spread and most of the information is available through the internet, not having access to computer means not being able to use the resources and information. Therefore, digital divide can also be seen in terms of difference in having or not having the opportunities to use and avail the resources, and recent developments. The difference in the opportunities between users and non-users of technology gives rise to one more issue of equity. Those with access of information and computing technologies (ICT) are the privileged people as they reap the benefit of technology. The non-users of ICT are lagging behind because they cannot reach the information and available resources. The digital equity can be seen between countries, gender, rural/urban areas. It is a general agreement among scholars that digital divide leads to several accumulated problems like poverty, illiteracy, unemployment, which makes digital divide a very serious problem and it requires lots of work to overcome it. The main object to overcome digital divide aims at reducing digital inequity i.e. unequal distribution of online resources and technologies. Digital equity focuses on equal distribution of technologies and resources among whole population. The digital equity can be seen among equal access of Information and communication technologies (ICT) in gender, age, countries, and rural/urban places.

2. Understanding the Digital Divide

This section addresses various matters related with the digital divide.

2.1 Is digital divide a Serious Issue?

In recent years, the use of computer is becoming a necessity of life. All important information is available online. In today's world, citizens are expected to be able to use and work on the computer and if they are not capable of it, then that leads to many problems associated with it [1]. The online access improves productivity and efficiency as it requires minimum effort to look and search for the information online. The growth of ICT is in every area of life and computers are used for social, personal, and educational purposes. The growth of these developments has created opportunities for everybody to have convenient and easy access of information. In education, multicultural education is becoming popular and whole world is getting connected with the use of technology. The use of computers is enabling sharing of educational resources among different countries. Online education is also rapidly increasing as it allows its users to take courses online at their convenience. Using online resources, different countries is different educational settings are able to benefit from each others experience. While sitting in one corner of the world, an online users can access information from anywhere in the world. The computers are successful in connecting the world together. The

ICT allows immediate access to news in every corner of the world, communication is becoming fast, and use of email lets people share each others ideas even if they are at distance. Though these developments are helping society to stay connected and progress, those without ICT are lagging behind in every aspect. The digital divide is serious if it continues because it will lead to divide the world into two parts: users/non-users of ICT. Those who do have access to technology are lagging behind in many areas where access to technology is needed. For instance, applying for jobs that are available only online, getting information about available resources, searching for available product, booking tickets to go somewhere, to explore the opportunities available in certain area. This list is endless. People with access to the recent developments and internet are in a better position to avail all these resources and information and benefit from them. This gives rise to digital divide as an equity issue too. The recent report by world information society [2] has indicated that the gap is widening and hence, measures are needed to minimize the gap.

2.2 Where is digital divide?

The digital divide can be viewed in terms of difference in the access to ICT in gender, countries, educated/illiterate, and rural/ urban area. In countries, the use of computers is extensive in some countries than other. Like developed countries have more percentage of people with access of ICT compared to developed countries. Even in developed countries there are certain regions that are behind others in term of access to ICT. Dijk's article [3] reports how southern regions in Europe are having limited access to ICT compared to rest of Europe. Among developed and developing countries, there are many reasons for the unequal distribution of ICT. In developing countries, the government is still struggling to provide education, power, food, and home—the issue of digital divide is not focused. The priority is to give importance to basic needs to life that the issue to digital divide is not completely addressed. In order to have access to ICT, developing countries need resources, necessary infrastructure and money. The report on world information society points out the recent developments in bringing ICT to the developing countries. In rural setting, access to ICT is not prevalent as urban area. The reason for easy access to ICT in urban area is due to more educated people, having power access, and better awareness of the benefits of ICT. In rural areas, people get limited access to ICT due to less availability of computers, learning opportunities, less resources, and less usage learning facilities. Among gender, the use of ICT is more prevalent in males compared to female. In some regions, females have less ICT learning opportunities compared to males. Females are bound by society to provide the basic necessities of life like food, cleaning, child-care, and other household work that they get less time to explore ICT. In technical education study, the percentage of males is more than females. This leads under-privileged people to get further behind their peers in access of ICT or gives rise to so called the digital divide.

2.3 How to Overcome digital divide?

The digital divide could be due to many causes or multiple factors. In order to combat digital divide, there is a need to understand the factors that leads to digital divide. Dijk [3] argues that access to ICT is possible in four steps: motivational, material, skill, and usage access. The author points out that only by having physical access; it is not possible to have access to ICT.

Motivational access is the first access needed for ICT and is described as not having access to ICT due to negative psychology about using technology. This view leads people to have bad views and fears about using ICT. Some people think that ICT are not useful for them as it won't benefit them in any way. Some people have natural fears that something bad could come if they use ICT—using computers with personal information may lead to identity theft. For instance, some people don't want to use computer because they think that it will make their lives public or don't want to use computers because they are not convinced about the security of the information or data while working online. In some situations, some people want to hide their limited knowledge of computers and it makes them staying away from using the computers. Age, gender, and social culture are some of the factors that could cause motivational access. Motivational access leads to digital divide since it works against working on computers. To overcome motivational access, a positive and conducive environment is needed. This could be done by increasing awareness about benefits of using the computers and providing many opportunities for them to learn using the computers.

Material access is the next access that comes after motivational access. This means having physical access to computers. So, owning a computer or having access to computers comes in this category. This access is increasing these days as computers are available for people to work at public places, number of people that own a computer are increasing, and access to computers is also possible by paying a minimal amount (using cyber cafes to work on computers).

After having motivation to use computers and having physical access to computer, next access needed is skill access. The user should have the skills to work on the computers. Skills access can be seen in terms of three categories: operational skills - knowing how to use internet for basic operations like working on internet, information skill - using computer to search for information on a particular area, and strategic skill - using computers for some specific purpose or goal. This includes knowledge of many computer software and programs. For instance, using computer to search for best books on calculus comes in skill access. Strategic skills require users to have net-working capabilities as well as knowledge of the area of interest.

Final access is usage skill that means using the computers. It comes after having motivation, physical access, and skills to use computers. Usage skill depends on many factors like usage time, usage application, Broadband (BB). Some people don't work on computers because of time limitation. Researchers argue about the increase in BB usage in recent years and use of BB in overcoming digital divide. The use of BB is more prevalent in areas than dial-up. BB allows users to have fast and easy access to ICT. It is further an debatable issue what role BB plays in reducing digital divide as some studies have pointed out that there are many household with BB access but ICT usage is very low or limited to certain members of the family. Users are not able to have access to ICT if any of these four accesses is missing.

Dijk [3] argues about the shifts in the strategies to overcome digital divide. The digital divide now is not about physical or material access but it is about skill and usage access. Author points out four types of access: motivational, material, skill, and usage and then emphasizes that though material access is closing, skill access is growing. So, the problem is not about having a computer but about using it. The strategies for overcome digital divide could be at large scale (all-over the countries) or small scale (directed towards a region). The article on digital divide in Africa refers to large scale strategy for digital divide that was focused on several countries on Java revolution whereas the article on digital divide in India focuses on the importance of small successful samples to successfully conquer digital divide. The small sample study show how to see if a certain approach might be successful in overcoming digital divide.

Some of the strategies that could stimulate ICT growth include - regulations by the government, market reforms, basic infrastructure, conducive environment for investment, and highly skilled labor. I think that all of these play a vital role in ICT growth and which strategy would be more effective than other varies from one region to other. Like when we see disparity in network assess from urban to rural area, government could be the one that plays most important role. By having rules and regulations this disparity could be minimized. I believe that in other strategies too like in basic infrastructure and market reforms, government plays an important role.

3. ICT and Education

The use of technology in education has increased in the last decade. Useful information and resources that could help learning are easily available on internet. For instance, the online availability of library resources makes learning possible at home. The articles, books, and resources are available for use, without physically going to the library. This not only saves time but also motivates users for education. This online availability of information helps both teachers and students. The teachers can find out about all useful books on any topic, look at the syllabus of other faculties teaching similar courses and use their inputs, design a lesson plan based on vastly available materials. Students can look for any kind of information on internet that could help them in clearing their doubt, use online available books and articles to sharpen their understanding of a subject, take part in online forums and discusses to broaden their understanding of a concept. The communication between faculty and student is also fast because of internet. The research also supports that technology could support teaching and learning ([4], [5]).

Some of the barriers in using ICT in teaching and learning are: lack of computing equipment, lack of institutional support, disbelief of technology values and benefits, lack of personal confidence in technology, and lack of time. Disbelief in technology values and benefits and lack of personal confidence in technology, can be grouped together in terms of motivational access. If the teacher does not want to use technology in the instruction this leads to digital divide. Also, the availability of technical equipment and facilities does not ensure that teachers' are going to use it if they lack personal confidence in the technology. To overcome these barriers a proper training for using technology should be given to the faculty. These two factors can be classified as internal factors that cause barrier to use technology in education. Lack of computing equipment and lack of institutional support can be classified as external factors leading to digital divide. Sometimes even if faculty wants to introduce the technology in instruction because of lack of computing equipment it is not possible. Sometimes there is not enough support for faculty to use technology in education. The supporters of traditional approach to instruction still focus on having instruction using chalk and board instead of using computers. The face-to-face instruction has limitation of having access by only those who can physically attend the class. The online instruction does not require person to be physically present, so you can learn at home too. Those who are far away (in different country), can also benefit from online education. The flexibility is one of the advantages of having online instruction as the person can take the lecture at per convenience.

3.1 The Role of the Teachers

Teachers play an important role in the use of ICT in education. If teachers motivate their students to use computers in the class and base their instruction that promotes using ICT, they could facilitate the use of computers among their students. There is a general agreement among researchers that use of ICT in education could benefit from teachers who are proficient in it. Some factors that have impact on teachers use of ICT are: lack of computing equipment, lack of institutional support, disbelief of technology values and benefits, lack of personal confidence in technology, and lack of time. Researchers have indicated that lack of institutional support and lack of time are factors that impact the use of ICT [6]. Salinas ([7]) pointed about the role of the teachers in rural areas. Author argues that teachers' are the gatekeeper in rural areas. The use of computers in rural area is depended on teachers' willingness and knowledge of computers. This study further reflects the importance of training teachers to use ICT so that they could incorporate technology in the education.

Hence, there is digital divide and it hampers the growth of under-privileged people. Some work is done to overcome digital divide but still more needs to be done.

4. The Digital Divide in Health

Professionals and general public use internet to seek important health related information. Internet not only allows easy and fast access, it is also inexpensive. eHealth refers to the healthcare practices that are supported by electronic resources and communication. Online health clinics are getting popularities these days. The issue of health gives rise to double divide since economic and social disparities reinforced divide on those who can't afford to have computers or internet access. Specifically, eHealth literacy was defined as "the ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem" ([8]).

The use of online resources in healthcare is popular in:

a) Electronic health records. This allows digital storage of health information which enables easy and fast exchange of health information between patients, healthcare professionals, and pharmacists.
b) Telemedicine. This refers to online use of patients' relevant information without any need to travel physically. It makes it possible for specialists to look at patients' information using online tools. This includes looking at patients data online or as simple as discussing a case on phone with other medical professional.
c) Health knowledge management. Resources that facilitate health literacy including many physicians' resources like Medscape.
d) mHealth. Using mobile services to collect patient's health data.
e) Virtual healthcare teams. This consists of groups of healthcare professionals who collaborate and share health services using online tools.
f) Healthcare information system. Some of the things that come in this are: online appointment scheduling, doctor's schedule, and looking at pharmacy open hours.

4.1 Factors Affecting the Use of Internet for Health Purposes

Renarchy et al. ([9]) have pointed out that use of internet for health purposes is very limited for primary prevention mostly it is used for secondary prevention as a tool to enhance understanding of a health issue. In the study, authors conducted a survey in Paris metropolitan area in 2005 to find out the relation between access to internet and use of internet for health purposes. Many researchers have claimed the association of age, education level, ethnicity, and income level with the use of online health services. Generally, young people are inclined towards using internet. White and rich people are using more online health services compared to other races and poor. Mostly educated people are the one who use internet for various purposes including health. Renarchy et al. find out that among age, educational level, race, and income level, educational level was the most significant variable that influences the use of computers for health purposes. Women were more inclined to search for health information using internet than men. This could be because women are considered as care taker in the family and they take the responsibility of looking after in case of sickness in the family. Renarchy et al. found a new dimension that influences the use—social isolation. Those who are not well connected with the society are less likely to use computers. This could be related to the fact that most of the times people search for health information related to someone close to them in family/friends. Use of internet was prevalent in Paris metropolitan than nationally. Socio-economic characteristics that lead to less use of internet are: low income, no education, social isolation, unemployed, and foreign nationality.

Health literacy of individuals is closely related with their education level. Several studies examined the relation of education and health literacy and concluded that the health literacy was higher in those who are well educated as they were able to understand, comprehend, and follow the direction to attain health. Taking medical doses as directed by the physician or taking over-dose is also related with individual's education level. Internet serves the purpose of disseminating health information and could help in developing better understanding of health related issues. Health awareness affects individual decision of going for preventive health care decisions like: regular medical check-ups, pap-smears, and having awareness of symptoms and consequences of deadly diseases like AIDS. Health literacy promotes one's understanding to better health and internet is one of the medium that provides health information and those who could not reach this information, are at disadvantaged condition. Health literacy is also related with race, socio-economic status (SES), and age. Researchers have shown that chronic diseases are prevalent in minority population. Also, population from low SES is more affected by deadly diseases and this could trace back to not having access to all health care facilities. Internet provides health literacy and using internet people from low SES could reach health information available at low cost.

4.2 Integrative Model of E-health

Bodie and Dutta ([10]) gave Integrative model of ehealth that provides a deeper insight to understand eHealth. Model depicts how micro issues combine together and result is a macro issue. To promote eHealth, there is a need focus on the inner issues that impede eHealth. Model suggests one of the micro issues is motivation. If users have no motivation to look online and search for the related information, there is no way he/she will use internet for health purposes. Model connects eHealth literacy with computer literacy and health literacy and points out importance of both for eHealth. For instance, if someone has some knowledge about some health issue and wants to further explore the possibilities using internet, both health literacy and computer literacy are needed. Using computer literacy that user will search for the specific information and using health literacy user will be able to discard irrelevant information and will be able to judge the credibility of given information.

This model also emphasizes four aspects that are needed to engage successfully on internet for eHealth. The first aspect is the ability to obtain information and have motivation to do it. This needs basic knowledge to search for the information on internet and have some inspiration to do it. The second aspect is to understand and rate the quality of given online information. This requires some basic understanding of the knowledge users are looking for that enables them to decide whether to discard it or take it seriously. Since online information is not always coming from a health care professional, it is must to be able to judge the significance of given information. The third aspect is to have competence and confidence to utilize the health information. If the users are health literate, they are able to comprehend and use the information in appropriate ways. And final aspect is to use all three above in an appropriate way, to have positive impact on health and well-being. All these steps occur when user uses online information for health purposes.

4.3 Credibility of Health Information Online

The issue of digital divide in eHealth is not limited to resources and software. Even those with these are not convinced to use internet for health. I believe that issue of health is very personal and the use of internet and relying on the information varies from person to person and also it's different in different contexts. Researchers have indicated that the information on internet is generally coming from those who are not professional in this field and hence how good it is requires personal judgment. Since the seekers and providers of internet information are increasing, it is important to understand the credibility of online health information. This is more important as when health information is provided online it is difficult to find out: (a) who authored the document, (b) when was it updated, and (c) how reliable it is.

Since the use of online health information is very much dependent on how much credibility users have in it, it is important to understand the issue of credibility of online information. Some researchers have concluded that perceived credibility of online information is same as information on television and radio but newspapers are perceived to be most credible than online resource of information. Institutional sites are perceived as more credible compared to individual sites. Eastin ([11]) studied the relationship topic knowledge, source expertise, and apparent credibility of online health information. One-hundred and twenty five students from a northwestern university participated in the study. The participants were randomly divided into two groups and were instructed to look at the selected websites that contain information about and unknown disease and a known disease like HIV. Participants were asked to assess the credibility of websites using different measures. Results indicate that source of the information plays an important role in credibility. If the information is coming from a highly credible source, it was viewed as credible. Knowledge of content was a significant factor in deciding the credibility of information. If users had knowledge about the content, they viewed the information as credible compared to the content about which users had no idea.

4.4 Myths About Digital Divide in Health

[11] Stellefson et al. ([12]) examined the myths about digital divide in health. They argued that the use of internet for health purposes has increased over past years. Educated people are more inclined to use internet for health purposes. Health websites on internet are serving their own purpose of having more and more customers instead of being there for common welfare. Commercial websites are often used compared to scholarly websites for health literacy. Some of the issues in eHealth are:

(a) Accessibility. One of the major factors that affect use of internet for health purposes is accessibility. Interoperable, knowledge based network like National Health Information Institute (NHII) provides reliable health information online. This type of interdisciplinary partnership allows promotes health information at low cost to users. Health educators need to come up with this kind of interdisciplinary collaboration that facilitates health literacy.

(b) Enabling eHealth. Health educators need to reap the benefits of technology to increase health awareness in the public. Some researchers have argued that the use of technology is more effective if it is coupled with entertainment. How technology could be used to increase health awareness is still not completely known. There is a need to provide a conducive environment online that promotes the usability of internet for health.

(c) Consumer health informatics. This is a new field that aims at analyzing consumer's health needs and implementing/providing consumer health information. This field is emerging and it is not possible to judge how effective it would be. There is a need for collaboration in fields like informatics and health education that makes health information available without fragmentation in different parts.

5. Conclusion

The future of education and health lies is reaping technological benefits. The digital divide in education and health is a serious issue as it impacts the quality of life one could have. A lot needs to be done in the area of overcoming digital divide especially in health. In the twenty-first century, we envision a society that uses technology for all those purposes that promotes health and improves quality of life. These envision allow us to have a society free from digital divide.

References

[1] V. Deursen, & V. Dijk, "Improving digital skills for the use of online public information and services". Government Information Quarterly, 26, 2009.

[2] "Chapter 2. Bridging the digital divide", World Information Society Report 2007-beyond WSIS, International Telecommunication Union, United Nations Conference on Trade and Development. International Telecommunication Union (ITU), Geneva, 2007.

[3] V. Dijk, "The Digital Divide in Europe". The Handbook of Internet Politics, Rutledge, London and New York, 2008.

[4] P. C. Gorski, P. C., "Insisting on the digital equity: Reframing the dominant discourse on the multicultural education and technology". Urban Education, 44, 2009.

[5] N. Parvathamma, N., "Digital divide in India: Need for correcting urban bias". Information Technology and Libraries, 22, 2003.

[6] S. Al-Senaidi, L. Lin, J. Poirot, "Barriers to adopting technology for teaching and learning in Oman". Computer & Education, 53, 2009.

[7] J. Salinas, "Digital inclusion in Chile: Internet in rural schools". International Journal of Educational Development, 29, 2009.

[8] C. D. Norman, H. Skinner, "eHealth literacy: Essential skills for consumer health in a networked world [Electronic Version]". Journal of Medical Internet Research, 8, 2006.

[9] E. Renachy, I. Parizot, P. Cauvin, "Health information seeking on the internet: a double divide? Results from a representative survey in the Paris metropolitan area, France, 2005-2006". BMC public health, 8, 2008.

[10] G. Boodie, M. Dutta, "Understading health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide". Health market quarterly, 25, 2008.

[11] M. Eastin, "Credibility assessments of online health information: The effects of source expertise and knowledge of content". Journal of Computer Mediated Communication , 6, 4, 2001.

[12] M. L. Stellefson, E. H. Chaney, J. D.Chaney, "The digital divide in health education: Myth or reality?"American Journal of Health Education, *39*, 106-112, 2008.

Author Biography

First Author: Veena is a graduate student in the Ph.D. program in Mathematics education at the University of Illinois at Urbana-Champaign. I received masters in Applied Mathematics in 2008 from University of Urbana-Champaign.

Second Author: Vaibhav sharma is assistant professor finance at Winthrop Univerty, SC.

A New Narrow-Block Mode of Operation for Disk Encompression with Tweaked Block Chaining

Debasis Gountia[1] and Dipanwita Roy Chowdhury[2]
(Corresponding author: Debasis Gountia)

[1]Department of Computer Science & Application, College of Engineering & Technology, Bhubaneshwar, India
[2]Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, India
dgountia@gmail.com, drc@cse.iitkgp.ernet.in

Abstract: In this paper, a new Disk Encompression (i.e., encryption with compression) with Tweaked Block Chaining mode (DETBC) has been proposed. DETBC is a modified of XTS i.e., Xor-Encrypt-Xor based Tweaked Code Book mode with CipherText Stealing. Unlike XTS, DETBC is faster, memory saving and is better resistant to the attacks. DETBC is characterized by its high throughput compared to the current solutions and improve its diffusion properties.

Keywords: block ciphers, disk encryption, Galois Field multiplier GF (2^{128}), tweakable block ciphers.

1. Introduction

Data encryption has been used for individual precious documents for the security purpose in the past. With the advent of more powerful desktop processors in the last decade, the data throughput of ciphers surpassed that of hard disks. Hence, encryption is no longer a bottle neck and regular users become more interested in the topic of hard disk encryption.

In today's computing environment, there are many threats to the confidentiality of information stored on computers and other devices like USB or external hard drive. Device loss or theft, Malware which give unauthorized access are common threat against end user devices. To prevent the disclosure of sensitive data, the data needs to be secured. Disk encryption is usually used to protect the data on the disk by encrypting it. The whole disk is encrypted with a single/multiple key(s) and encryption/decryption are done on the fly, without user interference. The encryption is on the sector level, that means each sector should be encrypted separately.

There are so many block ciphers dedicated to this task like Bear, Lion, Beast and Mercy [5, 5, 12, 16]. Bear, Lion and Beast are considered to be slow, as they process the data through multiple passes and Mercy was broken in [20]. The current available narrow-block modes of operations that offer error propagation are subjected to manipulation attacks. A need for a new secure and fast mode of operation with less memory consumption, that offers error propagation, has demanded.

In this paper, we propose a new narrow-block disk encryption mode of operation with compression. We decided to build the Tweaked Block Chaining (TBC) mode using Xor-Encrypt-Xor (XEX) [23] to inherit from its security and high performance and use CBC like operations to gain the error propagation property. This design is XEX-based TBC with CipherText Stealing (CTS) rather than Tweaked Code Book mode (TCB) as in case of XTS (XEX-based TCB with CTS). This model includes a Galois Field multiplier GF (2^{128}) that can operate in any common field representations. This allows very efficient processing of consecutive blocks in a sector. To handle messages whose length is greater than 128-bit but not a multiple of 128-bit, a variant of CipherText Stealing will be used for tweaked block chaining. We named this mode Disk Encompression with Tweaked Block Chaining (DETBC).

In section 2, we present Encryption with compression, and the constraints facing in the disk encryption applications. In section 3, we present tweak calculation, efficient multiplication, and exponential. Section 4 describes the implementation of our proposed scheme. Section 5 shows the performance analysis of narrow-block modes of operations that offer error propagation. Finally, section 6 concludes the work with presenting open problem.

2. Disk Encryption

Hard disk encryption is usually used to encrypt all the data on the disk. The whole hard disk is encrypted with a single/multiple key(s) and encryption/ decryption are done on the fly, without user interference. The encryption is on the sector level that means each sector should be encrypted separately.

2.1 Encryption with Compression

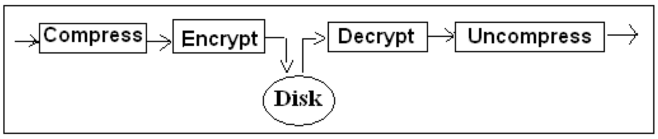


Figure 1. Steps for Disk encryption scheme.

Using a data compression algorithm together with an encryption algorithm makes sense for two reasons:
1. Cryptanalysis relies on exploiting redundancies in the plain text; compressing a file before encryption reduces redundancies.
2. Encryption is time-consuming; compressing a file before encryption speeds up the entire process.

In this work, we use the”LZW 15-bit variable Rate Encoder” [15] for compression of the data. To access data from the disk, we have to first decrypt and then uncompress the decrypted data.

2.2 Disk Encryption Constraints

The common existing disk constraints are:

Data size. The ciphertext length should be the same as the plaintext length. Here, we use the current standard (512-byte) for the plaintext.

Performance. The used mode of operation should be fast enough, as to be transparent to the users. If the mode of operation results in a significant and noticeable slowdown of the computer, there will be great user resistance to its deployment.

3. Disk Encompression with Tweaked Block Chaining

3.1 Goals

The goals of designing the Disk Encompression with Tweaked Block Chaining (DETBC) mode are:

Security: The constraints for disk encryption imply that the best achievable security is essentially what can be obtained by using ECB mode with a different key per block [21]. This is the aim.

Complexity: DETBC complexity should be at least as fast as the current available solutions.

Parallelization: DETBC should offer some kind of parallelization.

Error propagation: DETBC should propagate error to further blocks (this may be useful in some applications).

3.2 Terminologies

The following terminologies are used to describe DETBC:

P_i: The plaintext block i of size 128 bits.
J_s: The sequential number of the 512-byte sector s inside the track encoded as 5-bit unsigned integer.
I_i: The address of block i encoded as 64-bit unsigned integer.
T_i: The tweak i.
α: Primitive element of *GF* (2^{128}).
←: Assignment of a value to a variable.
‖: Concatenation operation.
PP_i : $P_i \oplus T_{i-1}$
K_1: Encryption key of size 128-bit used to encrypt the PP.
K_2: Tweak key of size 128-bit used to produce the tweak .
EK_1: Encryption using AES algorithm with key $K_{1.}$
DK_1: Decryption using AES algorithm with key $K_{1.}$
C_i: The ciphertext block i of size 128 bits.
$\oplus$: Bitwise Exclusive-OR operation.
$\otimes$: Multiplication of two polynomials in *GF* (2^{128}).

3.3 Tweak Calculation

In our proposed scheme, the mode of operation takes four inputs to calculate the ciphertext (4096-bit). These inputs are:

1. The plaintext of size 4096-bit.
2. Encryption key of size 128 or 256-bit.
3. Tweak key of size 128 or 256-bit.
4. Sector ID of size 64-bit.

Usually a block cipher accepts the plaintext and the encryption key to produce the ciphertext. Different modes of operation have introduced other inputs. Some of these modes use initial vectors IV like in CBC, CFB and OFB modes [7], counters like in CTR [8] or nonces like in OCB mode [9]. The idea of using a tweak was suggested in HPC [10] and used in Mercy [16]. The notion of a tweakable block cipher and its security definition was formalized by Liskov, Rivest and Wagner [11]. The idea behind the tweak is that it allows more flexibility in design of modes of operations of a block cipher. There are different methods to calculate tweak from the sector ID like ESSIV [13] and encrypted sector ID [14].

In this work, the term tweak is associated with any other inputs to the mode of operation with the exception of the encryption key and the plaintext. Here, an initial tweak T_0, which is equal to the product of encrypted block address, where the block address (after being padded with zeros) is encrypted using AES by the tweak key, and α^{Js}, where *Js* is the sequential number of the 512-byte sector s inside the track encoded as 5-bit unsigned integer and α is the primitive element of *GF* (2^{128}), will be used as the initialization vector (IV) of CBC. The successive tweaks are the product of encrypted block address and the previous cipher text instead of α^{Js}. When next sector comes into play, again initial tweak is used, and the successive tweaks are again the product of encrypted block address and previous ciphertext. This is done so assuming that each track has 17 sectors and each sector has 32 blocks as per the standard disk structure. This procedure continues till end of the input file.

3.4 Efficient Multiplication in GF (2^{128})

Efficient multiplication in *GF* (2^{128}) may be implemented in numerous ways, depending on whether the multiplication is hardware or software and optimization scheme. In this work, we perform 16-byte multiplication. Let a, and b are two 16-byte operands and we consider the 16-byte output. When these blocks are interpreted as binary polynomials of degree 127, the procedure computes p = a*b mod P, where P is a 128-degree polynomial $P_{128}(x) = x^{128} + x^7 + x^2 + x + 1$.Multiplication of two elements in the finite field *GF* (2^{128}) is computed by the polynomial multiplication and taking the remainder of the Euclidean division by the chosen irreducible polynomial. In this case, the irreducible polynomial is $P_{128}(x)= x^{128} + x^7 + x^2 + x + 1$.

Table 1. Algorithm for Multiplication in GF (2^{128}). Computes the value of p = a * b, where a, b and p ϵ GF (2^{128})

```
Algorithm PolyMult16( a, b) {
    p = 0;    /* Product initialized to zero*/
    while (b) {
        if (b & 1)  p = p ⊕ a; /*p xor a if the LSB of b is 1*/
        if (a127  = = 0) a << = 1; /*Left shift of bits in a by 1*/
        else a = (a <<= 1) ⊕ 0x87;/* x^128 + x^7 + x^2 + x + 1  */
        b >>= 1;/* Right shift of bits in the multiplier by 1 */
    }
    return p;
}
```

3.5 Efficient Modular Exponentiation

Compute efficiency: $z = x^c \bmod n$

Express c as follows: $c = \sum_{i=0}^{L-1} (c_i * 2^i)$,

where $c_i = 0$ or 1, value of i from 0 to (L-1) and L is the number of bits to represent c in binary.

The well-known Square-And-Multiply algorithm reduces the number of modular multiplications required to compute $x^c \bmod n$ to at most 2L. It follows that $x^c \bmod n$ can be computed in time $O(Lk^2)$. Total number of modular multiplications is at least L and at most 2L. Therefore, time complexity is the order of $[(\log c)*k^2]$, where n is a k-bit integer.

Efficient exponent in the finite field GF (2^{128}) is computed by the polynomial multiplication and taking the remainder of the Euclidean division by the chosen irreducible polynomial. In this case, the irreducible polynomial is $P_{128}(x) = x^{128}+x^7+x^2+x+1$.

Table 2. Algorithm for computing of $z = x^c \bmod n$, where x, c and $z \in GF(2^{128})$

```
Algorithm Square_And_Multiply ( x, c, n){
    z = 1; /* z initialized to one*/
    for (i = (L - 1); i > = 0; i--) {
        z = (z * z) mod n;
        if ( $c_i$ = = 1)
            z = (z * x) mod n;
    }
    return (z);
}
```

4. Implementation of the Proposed Scheme

The design includes the description of the DETBC transform in both encryption and decryption modes, as well as how it should be used for encryption of a sector with a length that is not an integral number of 128-bit blocks.

4.1 Encryption of a Data Unit.

The encryption procedure for a 128-bit block having index j is modeled with Equation (1):

$$C_i \leftarrow \text{DETBC-AES-blockEnc}(\text{Key}, P_i, I, j) \;\ldots\ldots\ldots\ldots(1)$$

where

Key is the 256-bit AES key
P_i is a block of 128 bits (i.e., the plaintext)
I is the address of 128-bit block inside the data unit
j is the logical position or index of the 128-bit block inside the sector
C_i is the block of 128 bits of ciphertext resulting from the operation

The key is parsed as a concatenation of two fields of equal size called Key_1 and Key_2 such that:

$$\text{Key} = \text{Key}_1 \,\|\, \text{Key}_2.$$

The plaintext data unit is partitioned into m blocks, as follows:

$$P = P_1 \,\|\, \ldots \,\|\, P_{m-1} \,\|\, P_m$$

where m is the largest integer such that 128(m-1) is no more than the bit-size of P, the first (m -1) blocks $P_1, \ldots, P_{m-1}$ are each exactly 128 bits long, and the last block P_m is between 0 and 127 bits long (P_m could be empty, i.e., 0 bits long).

The ciphertext C_i for the block having index j shall then be computed by the following or an equivalent sequence of steps (see Figure 2):

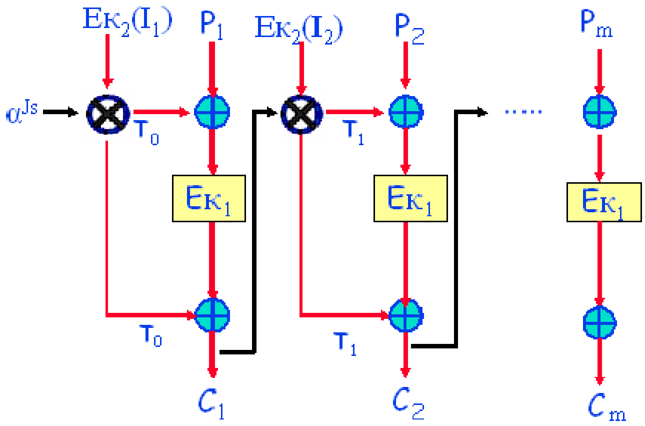


Figure 2. Encryption of data unit using DETBC.

```
Algorithm DETBC-AES-blockEnc(Key, $P_i$, $I_i$, j)
Case1 (j = 0):
    1. $T_{i-1} \leftarrow$ AES-enc ( $Key_2$, $I_i$ ) $\otimes\ \alpha^{Js}$
    2. $PP_i \leftarrow P_i \oplus T_{i-1}$
    3. $CC_i \leftarrow$ AES-enc( $Key_1$, $PP_i$)
    4. $C_i \leftarrow CC_i \oplus T_{i-1}$

Case2 (j > 0):
    1. $T_{i-1} \leftarrow$ AES-enc ( $Key_2$, $I_i$ ) $\otimes C_{i-1}$
    2. $PP_i \leftarrow P_i \oplus T_{i-1}$
    3. $CC_i \leftarrow$ AES-enc( $Key_1$, $PP_i$)
    4. $C_i \leftarrow CC_i \oplus T_{i-1}$
```

AES-enc(K, P) is the procedure of encrypting plaintext P using AES algorithm with key K, according to FIPS-197.The multiplication and computation of power in step (1) is executed in *GF* (2^{128}), where α is the primitive element defined in 3.2(see 3.4 & 3.5).

The cipher text C is then computed by the following or an equivalent sequence of steps:

```
Algorithm DETBC-Encrypt(Key, P, I )
    1. for i ← 0 to m-3 do
        a) j ← i % 32
        b) $C_{i+1} \leftarrow$ DETBC-AES-blockEnc (Key, $P_{i+1}$, $I_{i+1}$, j )
    2. r ← bit-size of $P_m$
    3. if r = 0 then do
        a) j ← (m-2) % 32
        b) $C_{m-1} \leftarrow$ DETBC-AES-blockEnc (Key, $P_{m-1}$, $I_{m-1}$, j)
        c) $C_m \leftarrow$ empty
```

```
4. else do
   a) $j \leftarrow (m-2) \% 32$
   b) $CC_{m-1} \leftarrow$ DETBC-AES-blockEnc(Key, $P_{m-1}$, $I_{m-1}$, j)
   c) $C_m \leftarrow$ first leftmost r bits of $CC_{m-1}$
   d) $C' \leftarrow$ last rightmost (128-r) bits of $CC_{m-1}$
   e) $PP_{m-1} \leftarrow P_m \| C'$
   f) $j \leftarrow (m-1) \% 32$
   g) $C_{m-1} \leftarrow$ DETBC-AES-blockEnc( Key, $PP_{m-1}$, $I_m$, j)
5. $C \leftarrow C_1 \| \ldots \| C_{m-1} \| C_m$
```

An illustration of encrypting the last two blocks $P_{m-1}P_m$ in the case that P_m is a partial block (r > 0) is provided in Figure 3.

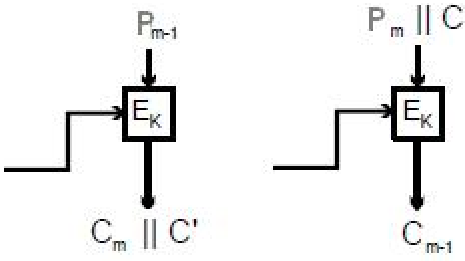


Figure 3. DETBC encryption of last two blocks when last block is 1 to 127 bits.

4.2 Decryption of a Data Unit.

The decryption procedure for a 128-bit block having index j is modeled with Equation (2):

$P_i \leftarrow$ DETBC-AES-blockDec(Key, C_i , I, j) ……….(2)

where

Key is the 256-bit AES key
C_i the 128-bit block of ciphertext
I is the address of the 128-bit block inside the data unit
j is the logical position or index of the 128-bit block inside the sector
P_i is the block of 128-bit of plaintext resulting from the operation

The key is parsed as a concatenation of two fields of equal size called Key_1 and Key_2 such that:

$Key = Key_1 \| Key_2$.

The ciphertext is first partitioned into m blocks, as follows:

$C = C_1 \| \ldots \| C_{m-1} \| C_m$

where m is the largest integer such that 128(m-1) is no more than the bit-size of C, the first (m-1) blocks $C_1,\ldots,C_{m-1}$ are each exactly 128 bits long, and the last block C_m is between 0 and 127 bits long (C_m could be empty, i.e., 0 bits long).

The plaintext P_i for the block having index j shall then be computed by the following or an equivalent sequence of steps (see Figure 4):

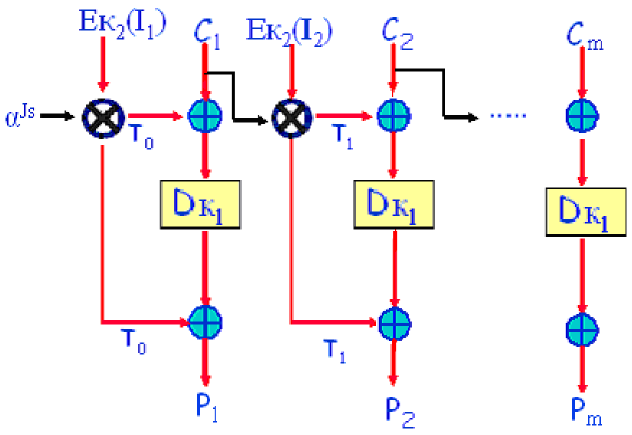


Figure 4. Decryption of ciphertext blocks using DETBC.

Algorithm DETBC-AES-blockDec(Key, C_i, I_i, j)

Case1 (j = 0):

```
1. $T_{i-1} \leftarrow$ AES-enc ( $Key_2$, $I_i$ ) $\otimes \alpha^{Js}$
2. $CC_i \leftarrow C_i \oplus T_{i-1}$
3. $PP_i \leftarrow$ AES-dec( $Key_1$, $CC_i$)
4. $P_i \leftarrow PP_i \oplus T_{i-1}$
```

Case2 (j > 0):

```
1. $T_{i-1} \leftarrow$ AES-enc ( $Key_2$, $I_i$ ) $\otimes C_{i-1}$
2. $CC_I \leftarrow C_i \oplus T_{i-1}$
3. $PP_i \leftarrow$ AES-dec( $Key_1$, $CC_i$)
4. $P_i \leftarrow PP_i \oplus T_{i-1}$
```

AES-dec (K,C) is the procedure of decrypting ciphertext C using AES algorithm with key K, according to FIPS-197. The multiplication and computation of power in step (1) is executed in *GF* (2^{128}), where α is the primitive element defined in 3.2 (see 3.4 & 3.5).

The plaintext P is then computed by the following or an equivalent sequence of steps:

Algorithm DETBC-Decrypt (Key, C, *I*)

```
1. for $i \leftarrow 0$ to m-3 do
   a) $j \leftarrow i \% 32$
   b) $P_{i+1} \leftarrow$ DETBC-AES-blockDec (Key, $C_{i+1}$, $I_{i+1,}$ j )
2. $r \leftarrow$ bit-size of $C_m$
3. if r = 0 then do
   a) $j \leftarrow (m-2) \% 32$
   b) $P_{m-1} \leftarrow$ DETBC-AES-blockDec (Key, $C_{m-1}$, $I_{m-1}$, j)
   c) $P_m \leftarrow$ empty
4. else do
   a) $j \leftarrow (m-1) \% 32$
   b) $PP_{m-1} \leftarrow$ DETBC-AES-blockDec(Key, $C_{m-1}$, $I_m$, j)
   c) $P_m \leftarrow$ first leftmost r bits of $PP_{m-1}$
   d) $C' \leftarrow$ last rightmost (128-r) bits of $PP_{m-1}$
   e) $CC_{m-1} \leftarrow C_m \| C'$
   f) $j \leftarrow (m-2) \% 32$
   g) $P_{m-1} \leftarrow$ DETBC-AES-blockDec(Key, $CC_{m-1}$, $I_{m-1}$, j)
5. $P \leftarrow P_1 \| \ldots \| P_{m-1} \| P_m$
```

An illustration of encrypting the last two blocks $C_{m-1}C_m$ in the case that C_m is a partial block (r > 0) is provided in Figure 5.

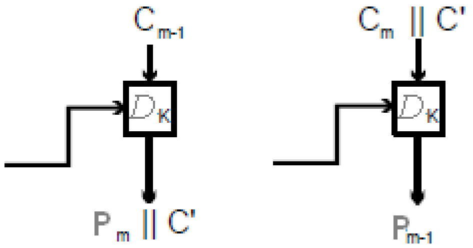


Figure 5. DETBC decryption of last two blocks when last block is 1 to 127 bits.

5. Performance Analysis

Security: Each block is encrypted with a different tweak T, which is the result of a non-linear function (multiplication) of encrypted file address and previous ciphertext (α^{Js} for 1st block); due to this step the value of the tweak is neither known nor controlled by the attacker. By introducing the tweak, the attacker can not perform the mix-and-match attack [21] among blocks of different sectors, as each sector has a unique secret tweak. Any difference between two tweaks result full diffusion in both the encryption and decryption directions. These enhance the security.

Here we also give option for the value of α to the user; it reduces the probability of getting plaintext from ciphertext. This is so because same plaintext produces different ciphertext if we choose different value for α. This also increases confusion.

Complexity: DETBC possesses high performance as it uses only simple and fast operations as standard simple shift and add (xor) operators are used in the multiplication in the finite field *GF* (2^{128}) having O(1) time complexity. Compression before encryption also enhances the speed and hence performance.

Parallelization: DETBC can be parallelized on the sector level as each sector is encrypted independently to other sectors. Also a plaintext can be recovered from just two adjacent blocks of ciphertext. As a consequence, decryption can be parallelized.

Error propagation: As each block depends on its previous block, a one-bit change in a plaintext affects all following ciphertext blocks. Hence, error propagation is met.

DETBC meets all its design goals.

5.1 Comparison

In this section, we compare our model with existing models [18]. The speed presented in table 3 for our mode (DETBC), is obtained from C implementation and taking a binary file as input, running on a 3 GHz Intel Pentium IV processor.

Table 3. Number of clock cycles reported by different mode of operation.

Mode	Key Length 128-bits
DETBC	4158
CBC	12630
CFB	12585
ESCC	12660

Note that the reported values are the minimum from measurements of different input files, to eliminate any initial overheads or cache misses factors. It is clear that DETBC possesses high throughput.

6. Conclusions and Open Problem

In this paper, we present a new mode of operation for disk encryption applications. The proposed scheme possesses a high throughput. Although, it is designed based on the CBC mode, it can be parallelized and does not suffer from the bit flipping attack. This mode also utilizes less memory space as the input file is first compressed and then it is encrypted.

There still remain many open problems in the search for efficient and secure disk encryption. There is a lack of good Boolean functions for the tweak generator which are efficient and also resist the cryptanalytic attacks, in particular algebraic and fast algebraic attacks.

References

[1] Bruice Schneier, Applied Cryptography, Wiley Press, Second Edition.

[2] Douglas R. Stinson, Cryptography Theory and Practice, CRC Press, Second Edition.

[3] Mark Nelson, Jean-Loup Gailly, The Data Compression Book, M&T Press, Second Edition.

[4] William Stallings, Cryptography and Network Security, Pearson Education, Fourth Edition.

[5] Anderson, R., Biham, E.: Two practical and provable secure block ciphers: BEAR and LION. In: Gollmann, D. (ed.) FSE 1996. LNCS, vol. 1039, pp. 113-120. Springer, Heidelberg (1996)

[6] S. Halevi and P. Rogaway, A tweakable enciphering mode, in Lecture Notes in Computer Science, D. Boneh, Ed. Berlin, Germany: Springer-Verlag, 2003, vol. 2729, pp. 482-499.

[7] A. Menezes, P. V. Oorschot., and S. Vanstone. Handbook of Applied Cryptography. CRC Press, 1996.

[8] D. McGrew. Counter Mode Security: Analysis and Recommendations. http:// citeseer.ist.psu.edu/mcgrew02counter.html, 2002.

[9] P. Rogaway, M. Bellare, and J. Black. OCB: A blockcipher mode of operation for efficient authenticated encryption. ACM Trans. Inf. Syst. Secur., 6(3):365-403, 2003.

[10] R. Schroeppel. The Hasty Pudding Cipher. The first AES conference, NIST, 1998. http://www.cs.arizona.edu/~rcs/hpc

[11] M. Liskov, R. L. Rivest, and D.Wagner, Tweakable block ciphers, in Lecture Notes in Computer Science, M. Yung, Ed. Berlin, Germany: Springer-Verlag, 2002, vol.2442, pp. 31-46.

[12] S. Lucks, BEAST: A fast block cipher for arbitrary block sizes. In: Horster, P. (ed.) Communications and Multimedia Security II, Proceedings of the IFIP TC6/TC11 International Conference on Communications and multimedia Security (1996)

[13] C. Fruhwirth, New Methods in Hard Disk Encryption. http://clemens.endorphin.org/nmihde/nmihde-A4-ds. pdf, 2005.

[14] N. Ferguson. AES-CBC + Elephant diffuser: A Disk Encryption Algorithm for Windows Vista. http://download.microsoft.com/download/0/2/3/0238acaf-d3bf-4a6d-b3d6-0a0be4bbb36e/BitLockerCipher200608.pdf,2006.

[15] Lempel-Ziv-Welch. http://en.wikipedia.org/wiki/Lempel-Ziv-Welch

[16] P. Crowley. Mercy, A fast large block cipher for disk sector encryption. In Bruce Schneier, editor, Fast Software Encryption: 7th International Workshop, FSE 2000, 2001.

[17] Mitsuru Matsui. The first experimental cryptanalysis of the data encryption standard. In Y. Desmedt, editor, Advances in Cryptology-CRYPTO 1994, number 839 in Lecture Notes in Computer Science, pages 1-11. Springer-Verlag, 1994.

[18] M. Abo El-Fotouh and K. Diepold, Extended Substitution Cipher Chaining Mode. http://eprint.iacr.org/2009/182.pdf

[19] P. Sarkar, Efficient Tweakable Enciphering Schemes from (Block-Wise) Universal Hash Functions. http://eprint.iacr.org/2008/004.pdf

[20] S. Fluhrer, Cryptanalysis of the Mercy block Cipher. In: Matsui, M. (ed.) FSE 2001. LNCS, vol. 2355, p. 28. Springer, Heidelberg (2002)

[21] I. P1619. IEEE standard for cryptographic protection of data on block oriented storage devices. IEEE Std. 1619-2007, April 2008. http://axelkenzo.ru/downloads/1619-2007-NIST-Submission.pdf

[22] P. Rogaway. Efficient Instantiations of Tweakable Block ciphers and Refinements to Modes OCB and PMAC. In Pil Joong Lee, editor, Advances in Cryptology - ASIACRYPT ’04, volume 3329 of LNCS, pages 16-31, 2004.

[23] Disk encryption theory. http://en.wikipedia.org/wiki/Disk_encryption_theory

[24] Latest SISWG and IEEE P1619 drafts for Tweakable Narrow-block Encryption. http://grouper.ieee.org/groups/1619/email/pdf00017.pdf

Debasis Gountia received the Bachelor of Computer Science and Engineering degree from Biju Patnaik University of Technology, Rourkela, India , in 2003. He received the Master of Technology degree in Computer Science and Engineering from the Indian Institute of Technology, Kharagpur, India in 2010.

Since January 2006, he has been a Lecturer with the College of Engineering & Technology, Bhubaneshwar, India. His research interests include cryptography, formal language and automata theory, operating systems, and distributed systems.

Dipanwita Roy Chowdhury received the Bachelor and the Master of Technology degree in Computer Science, both from University of Kolkata, India, in 1987 and 1989, respectively. She received the Ph.D. degree from the Indian Institute of Technology, Kharagpur, India in 1994.

She is a Professor with the Indian Institute of Technology, Kharagpur, India. Her research interests include cryptography, error correcting code, cellular automata, and VLSI design and testing.

Impact of Wormhole Attacks on MANETs

Saurabh Upadhyay [1] and Brijesh Kumar Chaurasia[2]

[1]SATI, Vidisha, INDIA,
[2] MIR Labs, Gwalior, INDIA,
Corresponding Adresses
[1]saurabh.cse.cs@gmail.com, [2]bkchaurasia.itm@gmail.com

***Abstract*:** A mobile ad hoc networks (MANETs) consists of a collection of wireless mobile nodes that are capable of communicating with each other. MANETs is infrastructure-less, lack of centralized monitoring and dynamic changing network topology. So, this network is highly vulnerable to attacks due to the open medium. In this paper, we discuss the impact of wormhole attack in MANETs. The wormhole attack is difficult to detect by using any cryptographic measures because they do not create any separate packets. In this work, several techniques of wormhole detection like watchdog, nodes with directional antenna and cluster based approach etc. Some prevention techniques such as packet leashes, time-of-flight, delphi protocol, pathrater technique etc. are also presented. The result analysis shows the impact of wormhole attack on MANETs in terms of throughput variations.

***Keywords*:** MANETs, Wormhole attack, Wormhole detection technique, Wormhole prevention, Attack model**.**

1. Introduction

A MANET is also known as a mobile mesh networks that consists of wireless mobile nodes that dynamically self organized connected by wireless links. Vehicular ad hoc networks and Sensor ad hoc networks are the varieties of MANETs.

In general, attacks are two types; active attacks and passive attacks. Wormhole attack [1] comes under active attack category is depicted in Fig. 1.

Passive attack: These types of attacks are not disrupting the network. For example eavesdropping attacks and traffic analysis and monitoring etc.

Active attacks: These types of attacks are disrupted the network, to alter or destroy data being exchanged in the network. These attacks can be internal or external.

Wormhole attack:

In wormhole attack [1], an attacker connects two distant points in the network, and then replays them into the network from that point. An example is shown in Fig. 2. Here *S* and *D* are the two end-points of the wormhole link (called as wormholes). In Fig. 2, wormhole attack is assumed between the node *A* and node *H* and their neighbor nodes, vice versa. The wormhole link can be established by many types such as by using ethernet cables, long-range wireless transmissions and an optical link in wired medium. Wormhole attack records packets at one end-point in the network and tunnels them to other end-point [2]. These attacks are severe threats to MANET routing protocols. For example, when a wormhole attack is used against an on-demand routing protocol such as AODV/ DSR, the attack could prevent the discovery of any routes other than through the wormhole.

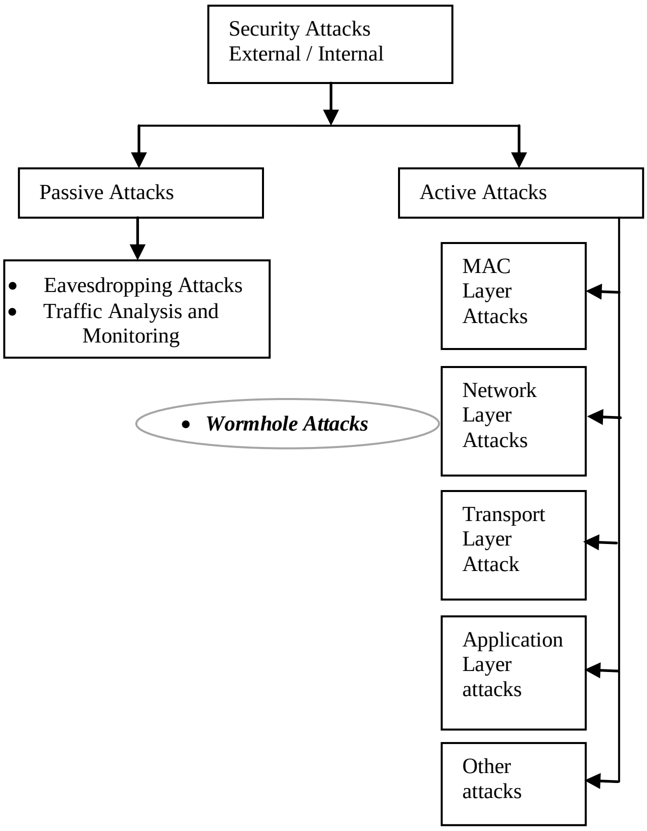


Figure 1. Categories of attacks in MANETs

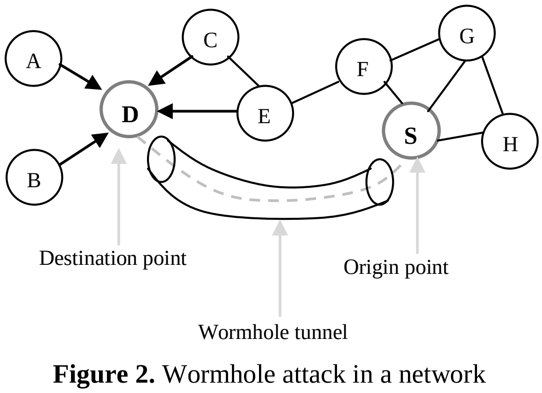


Figure 2. Wormhole attack in a network

The rest of the paper is organized as follows. Section 2 describes the problem. Section 3 of this paper presents the model and types of wormhole attack. In Section 4, we present the wormhole detection and prevention technique. Section 5 provides impact on MANETs. Section 6 concludes the work.

2. Problem Description

Wormhole attacks put severe threats to MANETs. This attack is very much dangerous because it can also still be performed even if the network communication provides authentication and confidentiality. Wormhole attack can also affect the network even if the attacker has no cryptographic keys. The wormhole attack is especially harmful against many ad-hoc routing protocols for example, ad hoc on-demand distance vector (AODV) [3], dynamic source routing (DSR) [4], the hop count of a path effects the choice of routes, clusterhead gateway switch routing protocol (CGSR) [5], hierarchical state routing protocol (HSR) [6] and adaptive routing using clusters (ARC) [7]. The wormhole attack is able to confuse the clustering procedure and lead to a wrong topology and it can partition the network through control links between two cluster heads of the routing hierarchy.

3. Wormhole Attack Model

A wormhole attack is consisting of two attackers and a tunnel through which the data is transmitted. For creating the wormhole attack the attacker creates a direct link referred as wormhole tunnel. The network which is caused by wormhole attack is depicted in Fig. 3.

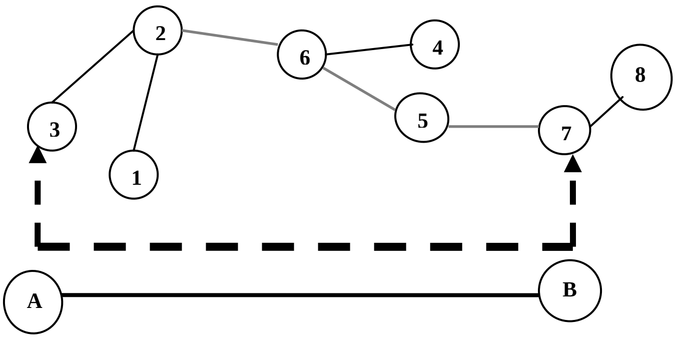


Figure 3. A netwok affected by wormhole

In Fig. 3, the tunnel represented by wired link, wireless out-of-bank link and logical link where the routing packet being encapsulated. When a wormhole tunnel has been created, attacker will receive packets from its neighbors and copies them and forwards them to the other attacker by using wormhole tunnel. Receiving node receive these tunneled packets. In a wormhole attack that uses wired links, high quality wireless out-of-band links, the attackers are directly connected to each other, so they can communicate very easily. However they require some special hardware to support such types of communication. A wormhole designed by using packet encapsulation is relatively much slower, but it can be launched very easily because it does not need any special hardware or special routing protocols.

Intruders A and B are connected by a wireless link, wireless link will be used to tunnel network data from the one end of the network to the other end of the network. Without presence of wormhole, node 7 and node 3 are apart from the cluster and their messages will forward to each node via nodes 2, 6 and 5. When wormhole attack is activated by intruders A and B, the node7 and node 3 are able to directly communicate to each others' messages and they will response that they are immediate neighbors.

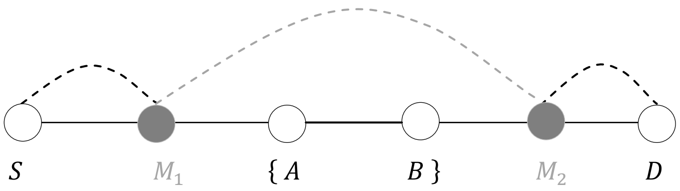


Figure 4a. Open wormhole

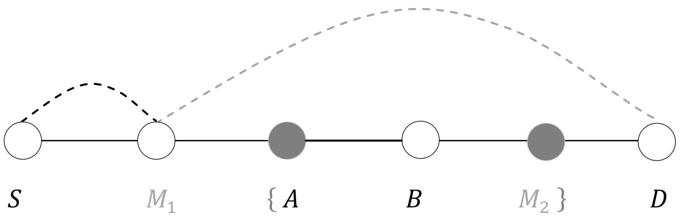


Figure 4b. Half open wormhole

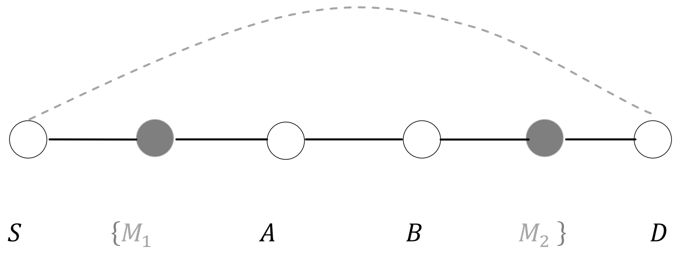


Figure 4b. Closed wormhole

If this happens, all communications between nodes 3 and 7 will be done by using the wormhole link introduced by A and B between node 3 and 7. The wormhole attack can be divided in three categories [8]; Open wormhole, half open wormhole and close wormhole.

In the given Fig. 4a, Fig. 4b and Fig. 4c M_1 and M_2 are presented the malicious nodes. S and D are the good nodes that are representing source and destination respectively. A and B are the good nodes between source and destination. The nodes in the curly-braces { } are the nodes that are on the path but are invisible due to presence wormhole and the curly-braces is presented here as false route in Fig. 4. Hence node S and D are connected by using a wormhole, so source and destination nodes think that they are immediate neighbors and all data between them will be transmitted by using this wormhole link. Both the nodes M_1 and M_2 are in the wormhole. In Fig. 4.b, M_1 node is the neighbor of source node S and it tunnels to destination through node M_2 and only one node can be seen by S and D due to wormhole attack. In the open wormhole attack both nodes M_1 and M_2 are visible to source node and destination node as shown in Fig. 4.a.

There is another classification of wormhole is discussed in [8], [10]. This classification is also categorized in three types;

i. Threshold based wormhole attack: In this category, wormhole will drop the packets of size greater than or equal to the threshold value.

ii. All pass based wormhole attack: In this type, wormhole will passes all packets irrespective of their size.

iii. All drop based wormhole attack: In this category, wormhole will drop all packets irrespective of their size.

4. Wormhole Detection and Prevention Techniques

In this section, we introduce the mechanism for detecting the wormhole attacks. To identifies misbehaving nodes and avoids routing through theses nodes, watchdog and pathrater is proposed in [11]. In this technique, watchdog identifies misbehavior of nodes by copying packets and maintained a buffer for recently sent packets. The overheard packet is compared with the sent packet, if there is a match then discards that packet. If the packet is timeout, increment the failure tally for the node. And if the tally exceeds the thresholds, then node will misbehave. The implementation of watchdog technique is shown in Fig. 5.

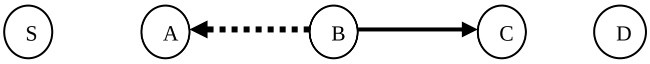


Figure 5. Watchdog implementation

In this figure, it is assumed that bidirectional communication symmetry on every link between nodes that want to communicate. If a node B can receive a message from a node A at time t, then node A could instead have received a message from node B at the time t will implement the watchdog. It maintain a buffer of recently sent packets and compares each overheard packet with the packet in the buffer, when B forwards a packet from S to D with the help of C, A can overhear $B's$ transmission and capable of verifying that B has attempted to pass the packet towards C. But this approach has some limitations and it is not detect the misbehaving node during ambiguous collisions, receiver collisions, false misbehavior and collusion.

The approach is used directional antenna to detect and prevent the wormhole attack [12]. The technique is assumed that nodes maintain accurate sets of their neighbors. So, an attacker cannot execute a wormhole attack if the wormhole transmitter is recognized as a false neighbor and its messages are ignored. To estimate the direction of received signal and angle of arrival of a signal it uses directional antennas. This scheme works only if two nodes are communicating with each other, they receive signal at opposite angle. But this scheme is failed only if the attacker placed wormholes residing between two directional antennas.

Statistical analysis scheme [13] is based on relative frequency of each link which is part of the wormhole tunnel and that is appears in the set of all obtained routes. In this techniques, it is possible to detect unusual route selection frequency by using statistical analysis detected and will be used in identifying wormhole links. This method do not requires any special hardware or any changes in existing routing protocols. It does not require even the aggregation of any special information, since it uses routing data that is already available to a node the main idea behind this approach resides in the fact that the relative frequency of any link that is part of the wormhole tunnel, will be much higher than other normal links.

In [14] is discussed graph theoretic model that can characterize the wormhole attack and can ascertain the necessary and sufficient conditions for the candidate solution to prevent wormhole attack. This scheme is also discussed a cryptographic based solution through local broadcast key and to set up a secure wireless ad hoc network against wormhole attacks. In this scheme, there are two types of nodes in the network named as: guards and regular nodes. Guards access uses GPS to access the location information or other localization method like secure range independent localization for wireless sensor network is presented in [15] and rebroadcast location data. Regular nodes need to calculate their location relative to the guards' beacons, thus they are able to distinguish abnormal transmission due to beacon retransmission done through the wormhole attackers. In this scheme, sender is encrypted all transmissions from local broadcast key and these information must be decrypted at the receiver end. But this scheme will be suffer the time delay to accumulate per node traveled and special localization equipment is needed to guard nodes for detecting positions.

To mitigate the wormhole attack in mobile ad hoc network, cluster based technique is proposed in [16]. In this approach clusters are formed to detect the wormhole attack. The whole network is divided into clusters. These clusters can either be overlapped or disjoint. Member nodes of cluster pass the information to the cluster head and cluster head is elected dynamically. This cluster heads maintains the routing information and sends aggregated information to all members within cluster. In this scheme, there is a node at the intersection of two clusters named as guard node. The guard node has equipped with power to monitor the activity of any node and guard the cluster from possible attack. The network is also divided into outer layer and inner layer. The cluster head of outer layer is having the responsibility of informing all nodes of the inner layer about the presence of the malicious node.

To prevent and detect the wormhole attack most common approach is discussed in [1] and [17], known as packet leashes mechanism. In this paper, they are presented two types of leashes: geographic leashes and temporal leashes also presented an authentication protocol. The authentication protocol is named as TESLA [18] with instant key disclosure and this protocol, for use with temporal leashes. In, geographic leashes each node access GPS information and based on loose clock synchronization. Whereas temporal leashes require much tighter clock synchronization (in the order of nanoseconds), but do not tightly depend on GPS information and temporal leashes that are implemented with a packet expiration time. The observation of this scheme is geographic leashes are less efficient than temporal leashes, due to broadcast authentication, where precise time synchronization is not easily achievable.

Other temporal leashes wormhole prevention technique is discussed in [19] based on time of flight of individual packets. This scheme is to measure round-trip travel time with its acknowledgment. This technique is used merkle hash tree and hash chains as explained in TESLA.

An efficient detection method known as delay per hop indication (DelPHI) for wormhole attack prevention is discussed in [20]. The protocol is developed for hidden wormhole attack and exposed wormhole attack. In this scheme, sender will check whether there are any types of malicious nodes presented in the routing path by that they will receive and implement the wormhole attacks. This scheme will not require clock synchronization, position information of nodes and any special types of hardwares.

Pathrater technique [11] calculates path metric for every path. By keeping the ratings of each node in the network, the path metric is calculated by using the node rating and connection reliability which is obtained from previous experience. Once the path metric has been calculated for all accessible paths, Pathrater will select the path with the highest metric. The path metrics would enable the Pathrater to select the shortest path. Thus it avoids routes that may have misbehaving nodes.

5. Impact of Wormhole Attack on MANETs

The wormhole attack is dangerous against the security in MANETs in which the nodes that hear a packet transmission directly from some node consider themselves to be in range of (and thus a neighbor of) that node. It is one of the most the powerful attack that are faced by many ad hoc network routing protocols. Since The wormhole attack does not require exploiting the feature of nodes in the network and it can interfere while executing the routing process. Attacker uses these attacks to gain unauthorized access to compromise systems or perform denial-of-service (DoS) attacks. In wormhole, the attacker at one end records the incoming traffic and tunnels packets to the other end. If routing control messages like RREQ are tunneled, this will result in distorted routing tables in the network. If there exist fast transmission path between the two ends of the wormhole that may tunnel the data at higher speed than the normal mode of wireless multi-hop communication. Thus, they will attract more traffic from their neighbors. This will results in rushing attack. In Rushing attack, due to the presence of fast transmission path all the packet will start following that path and this will increase the Average Attack Success Rate. Wormhole attack can also act as the first stage attackers where they can lead to the denial-of-service attacks. In the second stage, this may compromise the security of the global network as that breaks confidentiality and integrity. The wormhole attack is very harmful to the security of network. Due to the placement of the wormhole in the network there will be significant breakdown in communication across a wireless network. A successful wormhole attack may be the reason of disruption and breakdown of a network. Proper balance between these two is necessary to prevent much consumption of resources.

6. Simulation and Results

In this section, the impact of wormhole attack on MANETs is presented through simulation using QualNet [21]. The setup is shown in Fig. 6. The throughput is estimated by running the simulation experiment for 50 nodes in 1500x1500 m^2 area. increased and on increasing the data rate then packet drop is also increased. Fig. 7. and Fig. 8. represents the total packet sent and received at 2 Mbps constant bit rate (CBR). Fig. 9 depict the total packet sent and received at 11 Mbps CBR. Fig. 10. show the packet drop at nodes before or after the wormhole attack implementation. The result presents the packet drop is increased when attack is implemented in between the source and destination nodes. The observation and analysis shows that when wormhole is deployed on a route than packet drop is increased while maximizing the data rate, then packet drop is also swells.

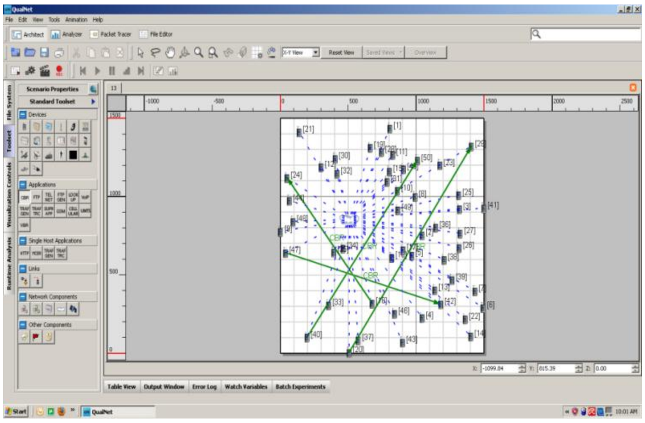


Figure 6. The simulation framework for wormhole attack

The simulation parameters are shown in Table1.

Table 1. Simulation elements

Simulation area	1500m x 1500m
Number of nodes	50
Physical Layer	802.11
MAC Layer	802.11b

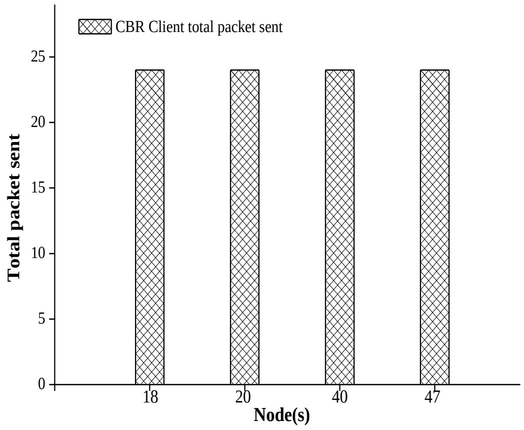


Figure 7. Total packet sent at 2 Mbps through client (CBR)

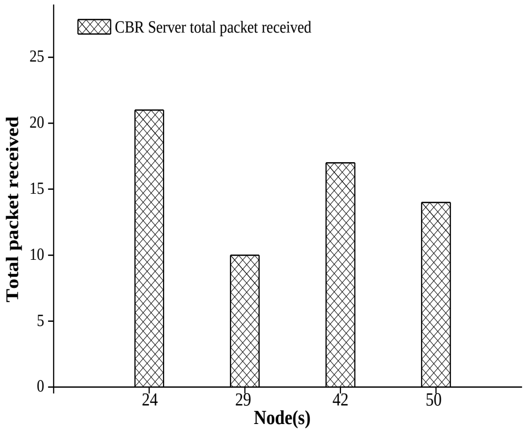


Figure 8. Total packet received at 2 Mbps

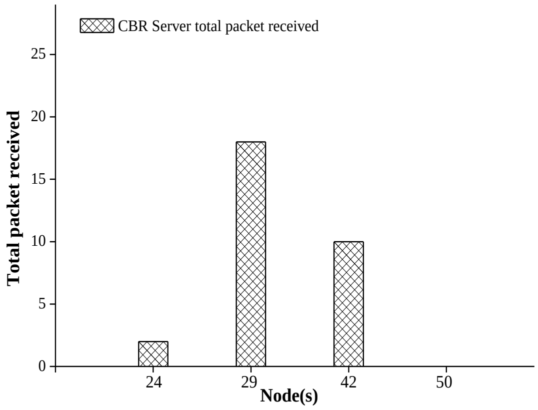


Figure 9. Total packet received at 11 Mbps

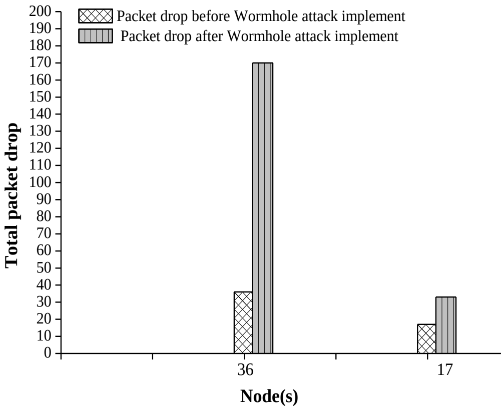


Figure 10. Packet drop at nodes

7. Conclusion

Wormhole attack is a very powerful attack that is created by malicious colluding nodes. It does not require any cryptographic breaks. The wormhole attack is a powerful attack that can have serious consequences on many proposed ad hoc network routing protocols. An attacker who can conducts a successful wormhole attack can disrupt routing, deny service to large segments of a network, creation of unconnected component within a network. In this paper we have discussed the several ways by which the wormhole can be handled. Results indicates that impact of wormhole attack is affected the throughput of packet ratio in terms of packet received, packet sent and packet drop at the nodes in ad-hoc networks as mobile ad hoc networks and sensor ad hoc networks. Future work on this topic will include developing any protocol that will prove much better security than existing against the wormhole attack.

References

[1] Y.-C. Hu, A. Perrig, D. B. Johnson, "Wormhole Attacks in Wireless Networks, Selected Areas of Communications," in IEEE Journal on, vol. 24, no. 2, pp.370-380, 2006.

[2] P. M. Jawandhiya, M. M. Ghonge, M. S. Ali and J.S. Deshpande, "A Survey of Mobile Ad Hoc Network Attacks ," in International Journal of Engineering Science and Technology, vol. 2, no. 9, 4063-4071, 2010.

[3] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in Proc. 2nd IEEE Workshop on Mobile Comput. Syst. Appl., pp. 90–100, 1999.

[4] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," in Proc. Conf. Commun. Architect., Protocols, Appl., pp. 234–244, 1994.

[5] Ching-Chuan Chiang, Hsiao-Kuang Wu, Winston Liu, and Mario Gerla, "Routing in clustered multihop, mobile wireless networks with fading channel," in Proceedings of IEEE Singapore International Conference on Networks (SICON '97), pp. 197–211, 1997.

[6] G. Pei, M. Gerla, X. Hong, and C.-C. Chiang, "A wireless hierarchical routing protocol with group mobility," in Proc. of IEEE on wireless communication and networking conference (WCNC), pp. 1538 - 1542, 1999.

[7] E. Royer, "Hierarchical routing in ad hoc mobile networks," Wireless Communication and Mobile Computing, vol. 2, no. 5, pp. 515-532, 2002.

[8] Khin Sandar Win, Pathein Gyi, "Analysis of Detecting Wormhole Attack in Wireless Networks," in World Academy of Science, Engineering and Technology 48, pp. 422-428, 2008.

[9] M. Jain and H. Kandwal, "A Survey on Complex Wormhole Attack in Wireless Ad Hoc Networks," in procedding of International conference on advances in computing, control and telecommunication technologies, pp. 555-558, 2009.

[10] S. Suresh Kumar, T.V.P. Sundararajan and Dr. A. Shanmugam, "Performance Comparison of Three Types of Wormhole Attack in Mobile Adhoc Networks," in proceedings of the Int. Conf. on Information Science and Applications ICISA, pp. 443-447, 2010.

[11] O. Kachirski and R. Guha, "Effective Intrusion Detection using Multiple Sensors in Wireless Ad hoc Networks", in Proc. 36th Annual Hawaii Int'l. Conf. on System Sciences (HICSS'03), pp.57.1, 2003.

[12] L. Hu and D. Evans, "Using Directional Antennas to Prevent Wormhole Attacks," in proceedings of the 11th Network and Distributed System Security Symposium, pp.131-141, 2003.

[13] N. Song, L. Qian and X. Li , "Wormhole Attack Detection in Wireless Ad Hoc Networks: a Statistical Analysis Approach," in proceeding of the 19th International Parallel and Distributed Processing Symposium (IPDPS'05), 2005.

[14] L. Lazos, R. Poovendran, C. Meadows, P. Syverson and L.W. Chang, "Preventing Wormhole Attacks on Wireless Ad Hoc Networks: A Graph Theoretic Approach," in IEEE WCNC 2005, Seattle, WA, USA, pp. 1193–1199, 2005.

[15] L. Lazos, and R. Poovendran, "SeRLoc: Secure Range-Independent Localization for Wireless Sensor Networks," in ACM WiSE'04, New York, NY, USA, pp. 73–100, October 2004.

[16] Debdutta Barman Roy, Rituparna Chaki and Nabendu Chaki, "New Cluster –based wormhole intrusion detection algorithm for mobile adhoc network," in International Journal of Network Security & Its Applications (IJNSA), vol. 1, no. 1, pp. 44-52, 2009.

[17] Y.-C. Hu, A. Perrig and D. B. Johnson, "Packet leashes: a defense against wormhole attacks in wireless networks," in Twenty-Second Annual Joint Conference of the IEEE Computer and Communication Societies, vol. 3, pp. 1976-1986, 2003.

[18] Adrian Perrig, Ran Canetti, Doug Tygar, and Dawn Song, "Efficient Authentication and Signature of Multicast Streams over Lossy Channels,"in Proceedings of the IEEE Symposium on Research in Security and Privacy, pp. 56–73, 2000.

[19] S. Capkun, L. Buttyan and J.-P. Hubaux, "SECTOR: Secure Tracking of Node Encounters in Multi-Hop Wireless Networks," in processings of the 1st ACM Workshop on Security of Ad Hoc and Sensor Networks, pp. 21-32, 2003.

[20] H.S. Chiu and K.S. Lui, "DelPHI: Wormhole Detection Mechanism for Ad Hoc Wireless Networks," in Proc. International Symposium on Wireless Pervasive computing, Phuket, Thailand, pp. 1-6, 2006.

[21] QualNet online available at: http://www.scalable-networks.com/products/system-requirements/qualnet/

Author Biographies

Brijesh Kumar Chaurasia is working in Privacy Preservation in Vehicular Ad hoc Networks. He is received his M. Tech. (Computer Science) degree from D.A.V.V., Indore, India. His research interest area is Security in Wireless Ad hoc Networks.

Saurabh Upadhyay is pursuing M. Tech. (Software System) degree from SATI, Vidisha, India. He is received his B. Tech. (Computer science) degree from GBTU (formally known as UPTU), India**.**
(saurabh.cse.cs@gmail.com)

Content Analysis of Mathematics Websites in Taiwan

Hsiu-fei Lee[1]
[1] Department of Special Education, National Taitung University, Taitung, Taiwan

fei@nttu.edu.tw

Abstract: Taiwan is well known for her products of PC and PC related accessories, wisdom mobile phones, and internet peripherals. With the advance of internet technology and raise of popularity rate of internet use world wide, besides the e-commerce opportunities, websites can also serve as a good platform to develop a country's cultural and creative industries, given the fact that Taiwan's international assessment on mathematics ranked number one based on the results of PISA in 2006. This study combined the content analysis method and four virtual spaces of ICDT model to analyze mathematics websites in Taiwan. The findings showed that most math websites in Taiwan remained at the stage of providing information, and were less on communication, and rarely on trade and delivery services.

Keywords: Content Analysis, Mathematics Websites, ICDT **.**

1. Introduction

Taiwan is well known for her products of PC and PC related accessories, wisdom mobile phones, and internet peripherals. With the advance of internet technology and the more affordable and portable tools, such as Eeepc, wisdom mobile phones, the number of internet users has dramatically raised in Taiwan and world wide in recent years. Below are some statistics reflecting such trend in Taiwan according to the government's reports. There are 14.46 millions of people above age 12 who are internet users in Taiwan; 70.9% of Taiwan's population use internet, and 80.7% of domestic households use internet in 2010 [1].

Taiwanese students have impressed the world for recent international assessments on mathematics. For instance, Taiwanese 15-year-old students ranked number one in mathematics in 2006 on Program for International Student Assessment (PISA), a test executed by the Organization for Economic Cooperation and Development (OECD). Given the facts above, it is interesting to learn how these two factors, in terms of mathematics and websites are integrated in Taiwan which plays the lead in the areas of PC technology and students' mathematics achievement. Thus, this study combined the content analysis method and four virtual spaces of ICDT model to analyze mathematics websites in Taiwan. In addition to the e-commerce opportunities, website design can serve as a good platform to develop a country's cultural and creative industries. The results of this study can provide the government, educational sectors, and the technology industries some advice for developing the soft power, in Taiwan and other countries as well.

2. Literature Review

2.1 Mathematics Websites

It is a trend to utilize website technology to study the learning effects of students in mathematics. There are 5 master's theses in Taiwan studying this subject matter and all found positive effects for students to learn mathematics by using the websites as a tool. Two studies investigated the junior high students, whereas the rest three studied the high school students [2-6]. The reasons were diverse, but mainly attributed to the merits of the design and technology of websites in terms of instant feedbacks, learner self-paced process, timeless constraints, vivid and diversified demonstrations and/or representations to boost the learners' motivation, attention, comprehension, and thus the learning effects lastly.

2.2 Content Analysis

Content analysis is a methodology which started in the field of social sciences for studying the content of communication. The sociologist, Earl Babbie [7] defined it as the study of recorded human communications, such as books, websites, paintings and laws. Ole Holsti [8] extended the use of such methodology beyond the original study in communication, and offered a broader definition of content analysis as a technique for making inferences by objectively and systematically identifying particular characteristics in the messages. Weber [9] defined content analysis as a methodology providing the procedures towards logical reasoning when analyzing texts or subject matters. Neuendorf [10] offered the most extensive definition for content analysis which included 6 parts. She stated in her book that "content analysis is a summarising, quantitative analysis of messages that relies on the scientific method (including attention to objectivity, inter-subjectivity, a priori design, reliability, validity, generalisability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented" (p. 10).

In summary, content analysis is a summarizing technique that relies on the scientific methods that can provide quantitative and qualitative information for the analyses. Thus, in this study the results of content analysis can help us understand the types and applications of technology for services, which thus can provide marketing strategies and decision makings for the development of cultural and creative industries.

2.3 Content Analysis of Websites

It has become popular to adopt content analysis as a method to analyze websites. There are many studies in Taiwan applied this method for various purposes which ranged from marketing, technology design to educational. The topics were quite diversified which included the comparisons of e-retail websites in Taiwan and China [11], floral e-commerce websites[12], digitized dimensions and layers of design of private museums, comparisons of leisure farms in Taiwan and China [13], travel websites [14] and educational websites [15-19].

However, the number of research on educational websites using content analysis method was comparatively scarce than other subjects, such as tourism or marketing. Only 5 studies in Taiwan investigating content analysis of educational websites of which subjects included elementary school websites[15], early childhood education[16], natural sciences [17], ideal English teaching [18], and mathematics [19]. There was only one study more related to content analysis of mathematics websites, nevertheless it did not directly analyze the mathematics websites. It focused on surveying the needs of teachers for an internet teaching resource center of mathematics. Therefore, it highlights the importance of this study—to analyze the content of mathematics websites in Taiwan.

2.4 ICDT

The ICDT model was developed by Albert Angehrn [20], which is a systematic approach to the analysis and classification of business-related Internet strategies. This model serves as a basis for identifying how existing products and services can be extended and/or redesigned in order to take advantage of the Internet, as well as suggesting how new goods and services become possible through this new medium.

This model makes the virtual market space into four areas. They are Virtual Information Space (VIS), Virtual Communication Space (VCS), Virtual Distribution Space (VDS), Virtual Transaction Space (VTS). Lueng [21] states that a firm's sustainability depends on the increase of its overall profitability, which can be achieved by increasing revenues or decreasing costs. The design and content of a firm's website can cause an impact on a firm's revenue change.

VIS offers the channels by which a firm can provide information about herself, the products, and the services. This area can allow global reach and the ability to provide rich information. VCS allows for a firm to exchange information with the various stakeholders in the business, i.e., suppliers, customers, and strategic allies. The information in the VIS is one way and more top-down, whereas communication in the VCS can be bi-directional and more mutual. The Internet has allowed for high-speed and low-cost communication, unhindered by physical and geographical constraints; e-mail, discussion groups, chat rooms, twitter are available and very convenient now. VTS is more than turning a firm's phone book into electronic version. It also includes ways of payment, security agreement, and customer services. VDS is a new distribution channel which a firm can quickly distribute the goods and services, especially those without a physical component, such as digitized media (e.g., books, music, software, games) and services (e.g., consulting, technical support, education).

The ICDT model can help analyze the current status of a firm's products and hence provide advice for future development and strategies to a firm.

3. Methodology

This study has reviewed the mathematics websites in Taiwan, compared the functions of current business websites, and then combined the results of the investigations with the ICDT model. The framework for the analysis of this study has been generated after synthesizing the data collected in the above mentioned procedures. As a result, the content framework of mathematics websites for the content analysis of this study is shown in Table 1.

3.1 Content Framework of Mathematics Websites

The VIS framework for mathematics websites combined the current functions of VIS with the concepts of *search* and *comparison* processes during the surfing procedure, and the *problem recognition* process during purchase decision procedure as well as the features of mathematics websites. As a result, it came up with this study's VIS which could be broken down into 4 parts. They included math website information, math website service, math teaching units, and teaching levels.

The VCS framework was based on the *examination* concept in the process of internet transactions. Examination means to examine or inspect the products that were on the shopping list of the customers. It consisted of two sections, members only and interaction. However, it requires the customers to register as members before they can read the basic information or to have any further inquiry of information on most mathematics websites.

The VTS framework was modified from the theory of three-stage internet transactions from the customers' perspectives, which were purchase determination, purchase consumption, and post-purchase interaction [22]. Yet, the procedures or stages were not necessarily linear; they could be coming back and forth. Many mathematics websites did not provide any commercial services, thus only commercial related websites were categorized in this section.

Lastly, the VDS framework included the last stage of the internet transactions which could be divided into two parts: delivery and return policy. The area of delivery was outdated with the advance of delivery business and competitions of global economy.

Table 1 Content Framework of Mathematics Websites

Virtual Space	Categorization Variables	
VIS	Math Website Information	News, activities, websites of math department, educational websites
	Math Website Services	Teaching materials, games, tests, teaching videos, journal articles/ reports, links, math articles
	Math Teaching Units	Number and quantity, geometry, algebra, statistics and probability, time
	Teaching Levels	K to elementary school, elementary school, junior high school, high school, college
VCS	Members only	Join member, contact us, newsletter, shopping cart info, membership

		value-added services, membership discount
	Interaction	Discussion board, message board, visitor counts
VTS	Products and Service Information	Cram schools, advertisement info, publishers, commercial online classes
	Type of Payment	Online credit payment, money transfer, TEL/FAX
	Security Agreement	Privacy policy, account and password security, copy rights
	Services	Return and exchange policy, Q&A services, membership services
VDS	Delivery	Logistics distribution, post office, pick-ups at convenience stores
	Return Policy	Pick-ups at home, post office, return by other delivery companies

3.2 Sampling

This study used 227 mathematics websites from Taiwan for final analysis. The source of data came from two biggest internet search engines in Taiwan—Yahoo and Google with keywords of mathematics (teaching) websites. The data collection time was from June 15 to November 15, 2010. There were 62 pages of results from Google and 100 pages from Yahoo which had been searched. On each page had 10 websites, and 609 results from Google and 997 results from Yahoo, that is, the total number of 1606 results were collected and searched. After deleting the results which were dead, unrelated to math websites, repetitive, or with risk alert from Google first, the same screening procedures were applied to the results from Yahoo. Consequently, a result of 900 internet links was gathered, from which consisted of 227 websites from Taiwan.

3.3 Reliability

Neuendorf [10] suggests that when human coders are used in content analysis, reliability translates to inter-coder/ inter-rater reliability. In other words, it means the amount of agreement or correspondence among two or more coders.

Wang [23] suggests the reliability coefficient to be 80% for the standards of the Gerbner's cultural index in the content analysis of communication. Thus, the reliability coefficients in the range between 67-80% are acceptable, of which results should be carefully defined and explained. In general, the reliability coefficient should be greater than 85% to show that the encoding results of the inter-coders are acceptable and reliable.

The formula for content analysis is as below:

$$\text{Intercoder agreement } = \frac{2M}{N_1+N_2}$$

M is the number of total agreement,
N_1 is the agreement number of the first encoder
N_2 is the agreement number of the second encoder

$$\text{Reliability} = \frac{n * \text{Average Intercoder Agreement}}{1+[(n-1)\text{Average Intercoder Agreement}]}$$

n is the number of encoders

There were two researchers who coded the websites based on the construct of the ICDT model. The results in Table 2 showed that the inter-rater agreement and reliability coefficient value of the Math websites in Taiwan was on average above .90. It means the results of the coding were reliable.

Table 2 Inter-rater Agreement and Reliability Coefficient Value of the Math Websites in Taiwan

Constructs Coefficient value Items	VIS	VCS	VTS	VDS
Inter-rater agreement	0.93	0.91	0.93	0.92
Reliability	0.96	0.95	0.96	0.96

4. Results

4.1 Analysis of VIS

The results of analysis of VIS showed that about 33.5% of Taiwan's math websites provided information for other educational websites. Math website service was the major function for Taiwan's math websites. Providing links to other resources came as the first (52.4%), and teaching materials (42.3%) as the second frequent function. Tests (32.2) and math articles (34.4) accounted for 1/3 of frequency. Games and teaching video were less frequent but still made up roughly 20 %. Journal articles and research reports was the least frequent. The reasons could be that journal articles were collected and linked to various data base. Therefore, the designer of the math websites in Taiwan did not provide such service. Such needs could be supplemented by the function of 'links' .

In other words, it seemed that the needs of getting teaching materials or information on education related issues, and/or even information on math articles could be more easily met than the other functions from the math websites in Taiwan. The rest two sections were related to teaching materials function. In addition, the math teaching units and levels were interrelated due to the system of national entrance examinations for high schools and colleges in Taiwan. The content of teaching units matched with the teaching levels in that these units were the core topics for the mathematics of the entrance examinations. As a result, it is not surprising to see that the percentage for the levels of junior high and high school was the same, 33.5%. The elementary school was a bit higher than junior high and high schools, 33.9%. This could be attributed to the fact that parents in Taiwan value education and consider math as an important subject for their children to be successful in moving up to higher education.

Time is a difficult topic for many children due to its abstract nature. We can measure time, but it is hard to really feel its existence. The trouble thing for children to learn time is the seemingly arbitrary units to be remembered and to convert. However, time as a teaching unit did not come as often on the math websites in Taiwan (11% only). This could be that time did not weigh as much as other topics for the examinations.

Surprisingly, the teaching level from kindergarten to elementary (usually below the 2nd grade) existed only 1.3%. Hopefully it had nothing to do with the entrance examinations.

Table 3 Statistics of VIS Categorization of Math Websites in Taiwan

Categorization variables \ Percentage		Taiwan (%)
Math Website Information	news	4.0%
	activities,	10.6%
	websites of math department	7.5%
	Educational websites	33.5%
Math Website Service	teaching materials	42.3%
	games	19.4%
	tests	32.2%
	teaching videos	19.4%
	journal articles/ reports	1.3%
	links	52.4%
	math articles	34.4%
Math Teaching Units	Number and quantity	23.8%
	geometry	26.0%
	algebra	20.7%
	statistics and probability	17.6%
	time	11.0%
Teaching Levels	K to elementary school	1.3%
	elementary school	33.9%
	junior high school	33.5%
	high school	33.5%
	college	16.7%

4.1 Analysis of VCS

The results of VCS showed that the functions of 'contact us' , 'join member' and 'message board' were the three most frequent functions on the math websites in Taiwan. However, the percentages did not exceed 20%. Shopping cart info, membership value-added services, membership discount, discussion board were less than 4%. The first three items were directly related to e-commerce and the low percentages suggested that this area of math websites in Taiwan can be further developed for VCS.

The differences between discussion board and message board were the directions of communication. Discussion board was more bi-directional; that is, back and forth discussions occurred in this section for the issues or questions posed. The level for the topics was usually more elaborated in discussion board. On the other hand, message board was usually for a quick or short question. Usually it was from parents to ask how to solve math problems or do math homework. The frequency for discussion board was only 1.8% when compared to 11.5% for message board.

Table 4 Statistics of VCS Categorization of Math Websites in Taiwan

Categorization variables \ Percentage		Taiwan (%)
Members only	join member	12.8%
	contact us	18.1%
	newsletter	4.8%
	shopping cart info	3.5%
	membership value-added services	1.8%
	membership discount	1.8%
Interaction	discussion board	1.8%
	message board	11.5%
	visitor counts	11.5%

4.2 Analysis of VTS

Generally speaking, the percentage for the analysis of VTS was on average below 10%. The results suggested that most math websites in Taiwan still did not provide e-commerce services. In fact, on average only less than 4% of the math websites in Taiwan engaged with e-commerce services. On the other hand, it was quite interesting to see that, comparatively to the low percentage for the functions in VTS, almost 10% of the math websites provided credit card or money transfer services.

Table 5 Statistics of VTS Categorization of Math Websites in Taiwan

Categorization variables \ Percentage		Taiwan (%)
Products and Service Information	cram schools	3.5%
	single ad info	1.3%
	publishers	0.4%
	commercial online classes	4.4%
Type of Payment	online credit payment	9.3%
	money transfer	9.3%
	TEL/FAX	7.9%
Security Agreement	privacy policy	4.4%
	account and password security	4.8%
	copy rights	5.3%
Customer Services	return and exchange policy	1.3%
	Q&A services	4.8%
	membership services	3.5%

4.3 Analysis of VDS

Due to the fact that the results of frequency of VTS were low due to not so many math websites in Taiwan engaged in e-commerce in this study, the results of analysis of VDS were not too different from those of VTS. In other words, when there was less trade, then there would be less delivery and return of products.

Table 6 Statistics of VDS Categorization of Math Websites in Taiwan

Categorization variable	Percentage	Taiwan (%)
Delivery	Logistics distribution	1.8%
	post office	1.3%
	pick-ups at convenience stores	1.3%
Return Policy	Pick-ups at home	0.9%
	post office	0.4%
	return by other delivery companies	1.3%

5. Conclusions

The operational functions of math websites in Taiwan were unevenly distributed in the four sections of the ICDT model. The order was VIS, VCS, DTS, and VDS. Generally speaking, the math websites in Taiwan still remained in the stage of providing information and then more traditional way of communication. Information for math teaching materials, tests, and links to other related websites took up the major functions for VIS. Teaching units and teaching levels were interrelated which were strongly affected by the national junior high and high school entrance examinations. Games and teaching videos were only nearly 20% of the functions. Only 1.3% of the function was for the level of kindergarten to elementary children. These three areas can be further developed for e-commerce, given the fact that education, particularly math education is what Taiwanese parents are willing to invest for their children. This is a cultural value that website designers, Internet Company should keep in mind.

The functions for VCS dropped to less than 20%. Some of them were even less than 2%, such as membership discount, membership value-added services, and discussion board. Furthermore, the results of VTS were even lower. Online math commercial websites only consisted of less than 5%. Consequently, it is needless to say the even lower number for the VDS because of low VTS percentage.

Overall, this corresponds with comments Lueng [21] made a decade ago. He stated that with the new internet technology, "this new communication channel can be used for lobbying, influencing opinions, negotiating potential collaborations, and the creation of communities. However, in most organizations, this is an undeveloped area".

References

[1] **Research, Development, and Evaluation Commission, ExecutiveYuan. http://www.rdec.gov.tw/ct.asp?xItem=4530943&ctNode=12232&mp=100 (12/14/2010)**

[2] Chen, Z. C. A case study of integrating website technology with mathematical teaching of second year of junior high school: Sum of exterior angles and interior angles, unpublished thesis, National Normal University, Kaohsiung, Taiwan, 2001.

[3] Sang, M. C. The influence of integrating "Changba's Mathland" website into mathematics instruction on learning achievement and attitudes toward mathematics. unpublished thesis, Fo Guang University, Yi-lang, Taiwan, 2005.

[4] Zhao, S.H. A study of using Moodle in high school mathematics for adaptive website learning, unpublished thesis, National Normal University, Kaohsiung, Taiwan, 2010.

[5] Wu, Y.S. A study on effects of web-based learning: A case of trigonometric function at senior high school level, unpublished thesis, National Normal University, Kaohsiung, Taiwan, 2004.

[6] Zhou, Z.M. A case study of internet applications for teaching mathematics to the freshmen in high school, unpublished thesis, National Normal University, Kaohsiung, Taiwan, 2003.

[7] Babbie, E. The practice of social research, 10th edition, Wadsworth: Thomson Learning Inc., 2003.

[8] Holsti, O.R. Content analysis for the social sciences and humanities reading, New York, NY: Addison-Wesley, 1969.

[9] Weber, R.P. Basic content analysis, 2nd ed., Newbury, CA: Sage Publications, 1990.

[10] Neuendorf, K.A. The content analysis guidebook, Thousand Oaks, CA: Sage Publications, 2002.

[11] Su, C.H. Analysis of Content framework of e-retail websites by cluster analysis and association rule technique, unpublished thesis, Kainan University, Tainan, Taiwan, 2005.

[12] Li, H.C., Chiang, J.Y., & He, C.H. Content analysis of Taiwan flower e-commerce websites, proceedings of the conference of technology and management association of Taiwan technology and management, Taipei, Taiwan, pp.235-243.

[13] Lin, C.N. & Shih, M.L. Content analysis on the website function structure of leisure farms across the Taiwan and mainland China, J. International Cooperation, 4(1), March 2009, pp. 16-30.

[14] Zhou, W.R. Content analysis of travel websites in Taiwan: Using the ICDT model, unpublished master's thesis, National Taipei College of Nursing, Taipei, Taiwan, 2004.

[15] Lu, L.D. Content analysis of elementary schools' websites, unpublished master's thesis, National Tainan University, Tainan, Taiwan, 2010.

[16] Hsu, H. C. Can you find it ? From the website needs for kindergarten teachers to explore early childhood education resource websites, unpublished master's thesis, National Taichung University of Education, Taichung, Taiwan, 2009.

[17] Huang, Y.P. A research on content concerning natural sciences on websites of elementary schools in changhua county, unpublished master's thesis, National Taichung University of Education, Taichung, Taiwan, 2007.

[18] Huang, Y.Z. A Study of the Ideal Elements of Domestic English Teaching Websites, unpublished master's thesis, National Taiwan Normal University, Taipei, Taiwan, 2008.

[19] Liao, Y.C. A study of establishing the internet teaching resource center of elementary school mathematics,

unpublished master's thesis, National Taipei Teachers College, Taipei, Taiwan, 2001.

[20] Angehrn, A. Designing mature internet business strategies: The ICDT model., European Management Journal, 15(4), pp. 361-369.

[21] Leung, A. The ICDT model: A framework for e-business, 1998. http://www.mediacircus.net/icdt.html **(12/1/2010)**

[22] Kalakota, R. & Whinston, A.B. Frontiers of electronic commerce. New York, NY: Addison-Wesley.

Author Biographies

Hsiu-fei Lee: Assistant professor, Department of Special Education, National Taitung University, Taitung, Taiwan. Ph.D of Communication Sciences and Disorders, specialized in learning disabilities, Northwestern University, U.S.A. MA of language, literacy, and culture, Stanford University, U.S.A. Research interests are mathematics learning disabilities, math low-achievers, math education of indigenous children, and culture and learning.

Anticipating Features of Ring Finger from Middle Finger Width: A Novel Method

Manimala.S[1] and C.N.Ravi Kumar[2]

[1]Department of CS & E, SJCE, Mysore, Karnataka, INDIA
[2] Department of CS & E, SJCE, Mysore, Karnataka, INDIA
Manimala.S
{malaharish,kumarcnr}@yahoo.com

***Abstract*:** The human hand is a masterpiece of mechanical complexity. Hands may be affected by many disorders, most commonly traumatic injury. In treating hand problems, the mastery of anatomy is fundamental in order to provide the best quality of care. The focus in this paper is on predicting geometric features of ring finger from the known width of the middle finger. Geometric features of both the hands from 100 people of different age group were extracted from the silhouettes. The proposed method can be used to predict ring finger length, position of knuckles and also finger width at the first and second knuckle using taalamana system and shilpa shastra. The estimation accuracy of more than 91% is achieved for all the estimated features of the ring finger.

***Keywords*:** ring finger features, golden mean, taalamana system, iconography, human hand, feature estimation**.**

1. Introduction

Human hand is the terminal part of the upper limb, used to manipulate or assess the environment. It is a highly mobile organ, capable of fine discriminative function and manipulation, both of which require a copious blood supply [26]. The anatomy of the hand is complex, intricate, and fascinating. Its integrity is absolutely essential for everyday functional living. Construction of the ring finger when only middle finger width is known is a challenging task. In view of this ring finger features are estimated using taalamana system and golden mean.

In case of accidents if only partial knowledge of the finger is available, then the proposed method can be used to obtain complete knowledge of the damaged part. In medical science when it is necessary to replace any part of the human body like fingers, it can be constructed using the features estimated by our proposed method for perfection in the plastic surgery.

1.1 Taalamana system

Iconography is the branch of art history which studies the identification, description, and the interpretation of the content of images. The word iconography literally means "image writing". The idea of constructing human hand is derived from Silpa Shastra. It has developed its own norms of measures and proportions. It is a complex system of iconography that defines rigid definitions [1,21,22]. The shilpa shastra normally employ divisions on a scale of one (eka tala) to ten (dasa tala). Each tala is subdivided into 12 angulas. It is called Taalamana paddathi or Taalamana system, the system of measurements by Tala, the palm of hand i.e. from the tip of the middle finger to the wrist as shown in figure 1.

1.2 Golden ratio

Two quantities are in the golden ratio if the ratio of the sum of the quantities to the larger quantity is equal to the ratio of the larger quantity to the smaller one. The golden section is a line segment divided according to the golden ratio. If a and b are the lengths of the larger and smaller line segments respectively, then golden ratio is represented as shown in equation 1.

$$\frac{a+b}{a} = \frac{a}{b} = \Phi(Phi) \qquad (1)$$

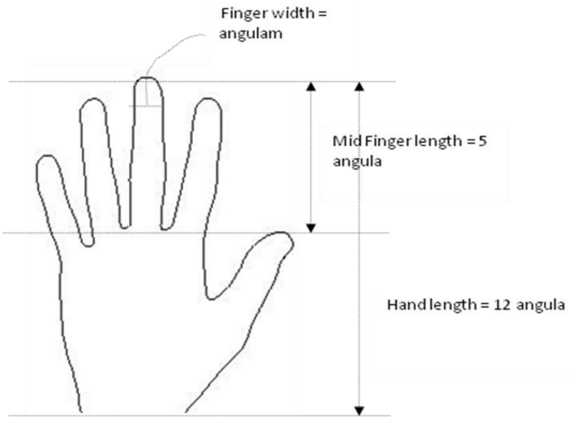


Figure 1: Computation of Middle finger length

The paper is organized into five sections. Introduction to taalamana system and golden ratio are given in first section. An insight into the related work is given in second section. Mathematical model is enumerated in section 3. In section 4 the proposed method is discussed and the simulation results are presented in section 5.

2. An Insight into the Related Work

Geometric measurements of the human hand have been used for identity authentication in a number of commercial systems. Anil K.Jain and others have worked extensively on hand geometry specifically for identification and verification systems [6,7,8]. There is not much open literature addressing the research issues underlying hand geometry-based identity authentication; much of the literature is in the form of patents [2, 3, 4]. Hand geometry recognition systems may provide three kinds of services like verification, classification and identification [12]. A novel contact-free biometric identification system based on geometrical features of the human hand is developed by Aythami Morales and others [11]. A component-based hand verification system using palm-finger segmentation and fusion was developed by Gholamreza and others. The geometry of each component of the hand is represented using high order Zernike moments which is computed using an efficient methodology [15].

Windy and others have used geometric measurements to study the sexual orientation. The ratio of the length of the second digit (2D) to the length of the fourth digit (4D) is greater in women than in men. This ratio is stable from 2 years of age in humans [9,10]. Gender classification from hand images in computer vision is attempted by Gholamreza and others [16].

Issac Cohen and others have worked on 3D hand construction from silhouettes of 2D hands [13]. Digital and metacarpal formulae are morphological variables which may also have functional significance in the understanding of how certain hand forms may be ill-fitted for certain tasks [14].

T.F.Cootes and others have worked on active shape models [17,18] which laid foundations for statistical shape analysis using Procrustes analysis, tangent space projection and Principal Component Analysis[19]. Geometric hand measurements are also used in hand gesture classification using a view-based approach for representation and Artificial Neural Network for classification [20].

3. Mathematical Model

Prediction of finger length, position of knuckles and finger width at the first and second knuckle of the ring finger are computed using taalamana system and golden ratio. The golden mean or ratio can be computed mathematically as shown in equation 2 and 3.

The middle finger length (MFL) is computed as five times the middle finger width (MFW1). Ring finger length (RFL) is computed using equation 4. Ring finger width (RFW1 and RFW2) are computed with the help of equation 5 and 6.

$$\frac{\sqrt{5}+1}{2} = \Phi(Phi) = 1.6180339 \qquad (2)$$

$$\frac{\sqrt{5}-1}{2} = \Phi(phi) = 0.6180339 \qquad (3)$$

Positions of the knuckles from finger tip and bottom of the ring finger (RL1 and RL2) are computed using the equation 7.

$$RFL = MFL - (phi * MFW1) \qquad (4)$$

$$RFW1 = MW1 - (\frac{MFW1}{16.0}) \qquad (5)$$

$$RFW2 = MW1 \qquad (6)$$

$$RL1 = RL2 = (phi * RFL) + (\frac{MFW1}{2.0}) \qquad (7)$$

4. Proposed Method

Silhouettes of both the hands of 100 users are taken. 24 features are extracted as shown in figure 2. For ring finger five features namely Ring Finger Width 1 (RFW1), Ring Finger Width 2 (RFW2), Ring Finger Length (RFL), Position of first knuckle from bottom(RL1) and position of second knuckle from finger tip (RL2) are extracted. Similarly for fore or index finger, middle finger and little finger these five features are collected and four features for the thumb totally to 24 feature set. From first width of the middle finger (MFW1), the values of RFL, RFW1, RFW2, RL1 and RL2 of ring finger are estimated. The actual and estimated values of a subset of samples are tabulated in table 1 and 2.

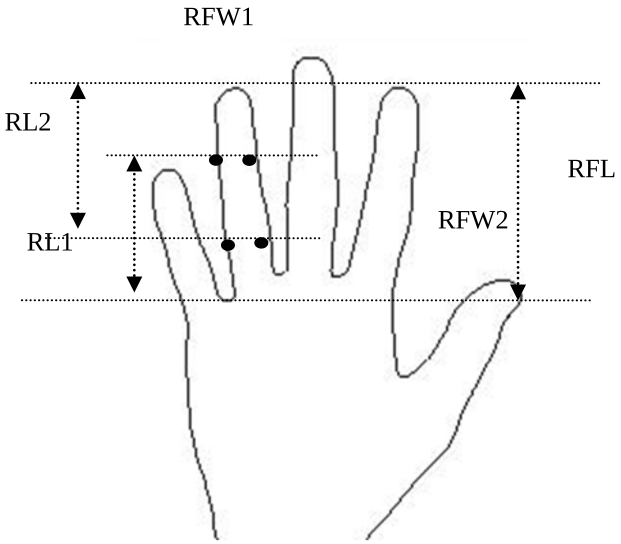


Figure 2: Feature Extraction of Ring Finger

5. Simulation Results

Geometrical features of both the hands are collected from 100 different people of different age group. Features collected for each of the finger are Finger Width (FW1, FW2), Finger Length (FL), Distance of first knuckle from bottom of the finger (L1) and distance of the second knuckle from the tip of the finger (L2). Total of 24 features are collected. In the current study ring finger features are estimated using only middle finger width.

In statistics, the mean square error or MSE of an estimator is one of ways to quantify the difference between an estimator and the true value of the quantity being estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. MSE measures the average of the square of the "error." The error is the amount by which the estimator differs from the quantity to be estimated. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate. The square root of MSE yields the root mean squared error or RMSE.

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error (MAE) is an average of the absolute errors computed as in equation 9, where f_i is the prediction and y_i the true value.

$$MSE = \frac{1}{n}\sum_{i=1}^{k}(f_i - y_i)^2 \qquad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{k} abs(f_i - y_i) \qquad (9)$$

Table 1 shows the actual and estimated values of RFL, RFW1 and RFW2 are shown. Absolute error and percentage of correctness for all the three features are also tabulated. Only 25 random samples are shown in the table. Similarly, in table 2 the actual and predicted positions of first knuckle (RL1) and second knuckle (RL2) of the ring finger are tabulated along with the absolute error and percentage of correctness. In table 3, the statistical features of the samples namely min, max, mean and standard deviation are tabulated. Table 4 shows RMSE, MAE and estimation accuracy for all the five features predicted for the ring finger using middle finger width. Mean absolute error and Root mean square error tabulated indicates that a maximum of 0.36 centimeters error is present in estimating position of the knuckles and approximately 0.5 centimeters in estimating ring finger length. Ring finger width shows an error of only 0.1 centimeters.

Table 1: Actual and predicted values of RFW1,RFW2 and RFL

MFW1	A-RFW1	P-R FW1	AE	%Corr	A-RFW2	P-RFW2	AE	%Corr	A-RFL	P-RFL	AE	%Corr
1.50	1.20	1.41	0.21	85.33	1.40	1.50	0.10	93.33	7.50	6.57	0.93	85.90
1.60	1.40	1.50	0.10	93.33	1.50	1.60	0.10	93.75	7.60	7.01	0.59	91.60
1.60	1.50	1.50	0.00	100.00	1.60	1.60	0.00	100.00	7.10	7.01	0.09	98.73
1.80	1.60	1.69	0.09	94.82	1.90	1.80	0.10	94.44	8.10	7.89	0.21	97.31
1.70	1.60	1.59	0.01	99.61	1.90	1.70	0.20	88.24	8.00	7.45	0.55	92.61
1.55	1.55	1.45	0.10	93.33	1.85	1.55	0.30	80.65	7.40	6.79	0.61	91.05
1.40	1.60	1.31	0.29	78.10	1.80	1.40	0.40	71.43	7.10	6.13	0.97	84.27
1.60	1.40	1.50	0.10	93.33	1.70	1.60	0.10	93.75	7.90	7.01	0.89	87.32
1.60	1.50	1.50	0.00	100.00	1.70	1.60	0.10	93.75	7.55	7.01	0.54	92.32
1.60	1.70	1.50	0.20	86.67	1.80	1.60	0.20	87.50	7.50	7.01	0.49	93.03
1.70	1.60	1.59	0.01	99.61	1.90	1.70	0.20	88.24	7.90	7.45	0.45	93.95
1.60	1.40	1.50	0.10	93.33	1.60	1.60	0.00	100.00	7.10	7.01	0.09	98.73
1.60	1.40	1.50	0.10	93.33	1.00	1.60	0.60	62.50	7.40	7.01	0.39	94.46
1.80	1.50	1.69	0.19	88.89	1.60	1.80	0.20	88.89	7.90	7.89	0.01	99.84
1.60	1.60	1.50	0.10	93.33	1.70	1.60	0.10	93.75	7.80	7.01	0.79	88.75
1.60	1.50	1.50	0.00	100.00	1.70	1.60	0.10	93.75	7.90	7.01	0.89	87.32
1.70	1.50	1.59	0.09	94.12	1.80	1.70	0.10	94.12	7.80	7.45	0.35	95.29
1.70	1.50	1.59	0.09	94.12	1.80	1.70	0.10	94.12	7.80	7.45	0.35	95.29
1.70	1.50	1.59	0.09	94.12	2.00	1.70	0.30	82.35	7.80	7.45	0.35	95.29
1.60	1.50	1.50	0.00	100.00	1.80	1.60	0.20	87.50	7.90	7.01	0.89	87.32
1.70	1.60	1.59	0.01	99.61	2.10	1.70	0.40	76.47	8.10	7.45	0.65	91.27
1.60	1.50	1.50	0.00	100.00	1.70	1.60	0.10	93.75	7.70	7.01	0.69	90.18
1.50	1.50	1.41	0.09	93.33	1.80	1.50	0.30	80.00	7.50	6.57	0.93	85.90
1.60	1.60	1.50	0.10	93.33	1.80	1.60	0.20	87.50	8.00	7.01	0.99	85.90
1.60	1.50	1.50	0.00	100.00	1.65	1.60	0.05	96.88	7.60	7.01	0.59	91.60

Table 2: Actual and predicted values of RL1 and RL2

MFW1	A-RL1	P-RL1	AE	%Correct	A-RL2	P-RL2	AE	%Correct
1.50	4.70	4.81	0.11	97.67	5.00	4.81	0.19	96.10
1.60	4.90	5.13	0.23	95.46	5.20	5.13	0.07	98.69
1.60	4.80	5.13	0.33	93.51	4.90	5.13	0.23	95.46
1.80	4.90	5.77	0.87	84.86	6.00	5.77	0.23	96.10
1.70	5.40	5.45	0.05	99.02	5.50	5.45	0.05	99.15
1.55	4.45	4.97	0.52	89.49	5.20	4.97	0.23	95.43
1.40	4.60	4.49	0.11	97.58	5.10	4.49	0.61	86.45
1.60	5.20	5.13	0.07	98.69	5.20	5.13	0.07	98.69
1.60	5.60	5.13	0.47	90.90	4.70	5.13	0.43	91.57
1.60	5.10	5.13	0.03	99.36	5.00	5.13	0.13	97.41
1.70	5.60	5.45	0.15	97.32	5.20	5.45	0.25	95.35
1.60	4.90	5.13	0.23	95.46	4.70	5.13	0.43	91.57
1.60	4.40	5.13	0.73	85.72	5.40	5.13	0.27	94.80
1.80	5.40	5.77	0.37	93.51	5.30	5.77	0.47	91.78
1.60	5.30	5.13	0.17	96.75	5.60	5.13	0.47	90.90
1.60	5.10	5.13	0.03	99.36	5.60	5.13	0.47	90.90
1.70	4.70	5.45	0.75	86.18	5.40	5.45	0.05	99.02
1.70	4.80	5.45	0.65	88.01	5.50	5.45	0.05	99.15
1.70	5.00	5.45	0.45	91.68	5.90	5.45	0.45	91.82
1.60	4.80	5.13	0.33	93.51	5.40	5.13	0.27	94.80
1.70	5.30	5.45	0.15	97.18	5.90	5.45	0.45	91.82
1.60	4.90	5.13	0.23	95.46	5.00	5.13	0.13	97.41
1.50	4.50	4.81	0.31	93.51	5.80	4.81	0.99	79.47
1.60	4.70	5.13	0.43	91.57	5.70	5.13	0.57	88.95
1.60	5.30	5.13	0.17	96.75	5.20	5.13	0.07	98.69

In figure 3(a-e) around 40 – 50 subset of the samples are plot indicating the actual and predicted values of RFL,RFW1, RFW2, RL1 and RL2 respectively.

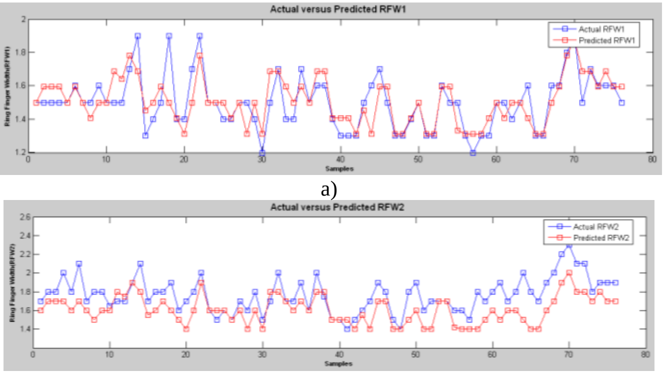
Red line in the plot shows the actual or true values and blue line indicates the predicted values. Overlapping in the graph shows the close relation of predicted values to the actual values.

Table 3 : Statistical Analysis

	Min	Max	Mean	Std Deviation
MFW1	1.3	2.0	1.5712	0.1246
RFL	5.90	9.20	7.39	0.57
RFW1	1.10	1.90	1.43	0.15
RFW2	1.00	2.30	1.65	0.20
RL1	3.80	6.10	4.92	0.46
RL2	4.10	6.50	5.03	0.45

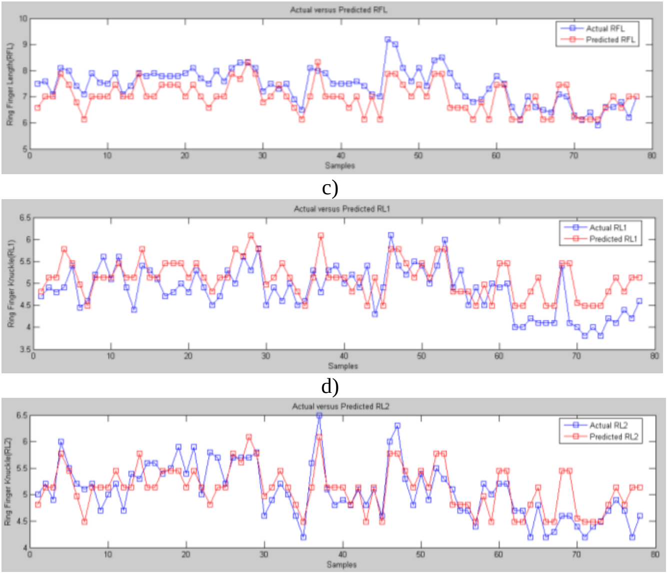
Table 4 : RMSE and MAE

	MAE	RMSE	Estimation Accuracy
RFL	0.56	0.66	91.68
RFW1	0.10	0.12	93.44
RFW2	0.14	0.18	91.17
RL1	0.36	0.45	92.86
RL2	0.27	0.34	94.73



a)

b)



c)

d)

e)

Figure 3(a-e) : Plot of actual and predicted values of RFW1, RFW2, RFL, RL1 and RL2

Conclusion

To the best of our knowledge this is the first humble beginning in estimating the geometrical features of ring finger from the width of the middle finger. In view of this Taalamana system and golden ratio are used to predict the feature values for RFL, RFW1, RFW2, RL1 and RL2. The graph in figure 3 indicates close association of the actual and the estimated feature values. Estimation accuracy of 92%, 93%, 91%, 93% and 95% for RFL, RFW1, RFW2, RL1 and RL2 features respectively is achieved. Only middle finger width is sufficient to estimate the features of the ring finger.

References

[1] Gopinatha Rao, T. A (1920). Talamana, or, Iconometry : Memoirs of the Archaeological Survey of India ; no. 3. Calcutta: Supt. Govt. Print

[2] R. P. Miller, “Finger dimension comparison identification system", US Patent No. 3576538, 1971.

[3] R. H. Ernst, “Hand ID system", US Patent No.3576537, 1971.

[4] H. Jacoby, A. J. Giordano, and W. H. Fioretti, “Personnel Identification Apparatus", US PatentNo. 3648240, 1972.

[5] Raul Sanchez –Reillo, Carmen Sanchez-Avila, Ana Gonzalez-Macros, “Biometric identification through hand geometric measurements”, IEEE Transactions on pattern analysis and machine intelligence, Vol 22, No. 10, Oct 2000.

[6] Anil K. Jain, Arun Ross, Sharath Pankanti, “A Prototype Hand Geometry-based Verification System”, 2nd International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), Washington D.C., pp.166-171, March 22-24, 1999.

[7] A.K. Jain, A. Ross, and S. Pankanti. A prototype hand geometrybased verification system. In Proceedings of 2nd Int’l Conference on Audio- and Video-based Biometric Person Authentication, pages 166–171, March 1999.

[8] Anil K. Jain and Nicolae Duta. Deformable matching of hand shapes for verification. In Proceedings of International Conference on Image Processing, October 1999.

[9] Windy M. Brown, Melissa Hines, Briony A. Fane, S. Marc Breedlove, “ Masculinized Finger Length Patterns in Human Males and Females with Congenital Adrenal Hyperplasia”, Hormones and Behavior 42, 380–386 2002, Elsevier Science (USA)

[10] Windy M. Brown, B.A., Christopher J. Finn, B.A., Bradley M. Cooke, and S. Marc Breedlove,“Differences in Finger Length Ratios Between Self-Identified Butch and Femme Lesbians”, Archives of Sexual Behavior, Vol. 31, No. 1, February 2002, pp. 123–127

[11] Aythami Morales, Miguel A. Ferrer, Francisco Díaz, Jesús B. Alonso, Carlos M. Travieso, “Contact-free hand biometric system for real environments”,16th European Signal Processing Conference, Lausanne, Switzerland, August 25-29, 2008.

[12] Yaroslav Bulatov, Sachin Jambawalikar, Piyush Kumar, Saurabh Sethia, “Hand recognition using geometric classifiers” Biometric Authentication, Lecture Notes in Computer Science, 2004, Volume 3072/2004, 1-29

[13] Isaac Cohen, Sung Uk Lee , “3D Hand and Fingers Reconstruction from Monocular View”, 17th International Conference on Pattern Recognition, 2004

[14] Stephen Lewis “Morphological aspects of male and female hands”, Annals of Human Biology,1996 Nov-Dec;23(6):491-4.

[15] Gholamreza Amayeh, George Bebis, Ali Erol, Mircea Nicolescu, “A Component-Based Approach to Hand Verification”, IEEE Conference on Computer Vision and Pattern Recognition, 2007

[16] Gholamreza Amayeh, George Bebis, Mircea Nicolescu, “Gender Classification from Hand Shape”, IEEE Computer Society Conference on Computer Vision and Pattern recognition workshops, 2008

[17] T. F. Cootes and Taylor, “Active shape models – smart snakes”, British Machine Vision Conference, pages 266–275, 1992.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models”, IEEE Transactions On Pattern Recognition and Machine Intelligence, 23(6):681–685, 2001.

[19] Mikkel B. Stegmann and David Delgado Gomez, “A Brief Introduction to Statistical Shape Analysis”, a data report at http://www.imm.dtu.dk/~mbs/

[20] Sanjay Kumar, Dinesh K Kumar, Arun Sharma, and Neil McLachlan “Classification of Hand Movements Using Motion Templates and Geometrical Based Moments”, IEEE, ICISIP 2004

[21] Gift Siromoney; M. Bagavandas, S.Govindaraju (1980). "An iconometric study of Pallava sculptures". Kalakshetra Quarterly 3 (2): 7–15.

[22] http://www.cmi.ac.in/gift/Iconometry/icon_pallavasculpture.htm

[23] Kramrisch, Stella; Raymond Burnier (1976). The Hindu Temple. Motilal Banarsidass Publ.. pp. 309. ISBN 9788120802247.

[24] Wangu, Madhu Bazaz. Images of Indian Goddesses: Myths, Meanings and Models. Abhinav Publications. pp. 72. ISBN 978817017416

[25] Manimala.S, Dr. C N Ravi Kumar, " Prediction of Middle Finger Features from its Width: A Novel Approach", International Journal of Advanced Research in Computer Science , Vol 1, No. 4, Nov-Dec 2010, pp 42-46, ISSN 0976 – 5697

[26] http://encyclopedia.stateuniversity.com/pages/9384/hand.html

Author Biographies

Ms. Manimala.S is a Senior Lecturer of the department of Computer Science and Engineering at Sri Jayachamarajendra College of Engineering Mysore. She obtained her Bachelor's degree in Computer Science and Engineering in 1994 and Masters in Software Engineering in 2004. Currently she is pursuing her Ph.D under the guidance of Dr.C.N. Ravi Kumar. She has five research papers published in International journal and National and International Conference. Her research interest includes Image Construction, Image Interpolation, Image Rendering, Pattern Recognition.

Dr. C.N. Ravi Kumar is the founder faculty of the department of Computer Science and Engineering at Sri Jayachamarajendra College of Engineering Mysore. He is presently working as a Professor and Head of the department. He obtained his Bachelor degree in Electronics and Communication in the year 1979,obtained his Master Degree in the year 1985 and was the first person to obtain the M.Sc. (Engg.) by Research degree from Mysore University. He obtained his doctoral degree in Computer Science and Engineering during the year 2000, under the guidance of Dr. K. Chidananda Gowda, the former Vice-Chancellor of Kuvempu University. He has 108 papers published to his credit in National, International journals and Conferences. His research interest includes Pattern Recognition, Image Processing, Biometrics, Data Mining.

Information Extraction from Remote Sensing Image (RSI) for a Coastal Environment Along a Selected Coastline of Tamilnadu

K. Bhuvaneswari[1], R. Dhamotharan[2], N. Radhakrishnan[3]

[1]Research Scholar, Mother Teresa Women's University, Kodaikanal, India
[2]Reader, Department of Botany, Presidency College, Chennai, India
[3]Director, Geocare Research Foundation, Chennai, India

Corresponding Author: radhakrishnan.nr@gmail.com

Abstract: With rapid development in spatial technology and with availability of tremendous amount of satellite data, studying and analyzing environment of an area especially along coast has become more meaningful. The synoptic view and repetitive coverage have paved way for such analysis under uniform illumination. Moreover the ability of such data in digital format and its nature has opened many unknown avenues to be explored in the arena of information extraction. At the same time, they have equally introduced certain complexities such as different digital values in different spectral region and varying spatial and radiometric parameters. Hence, applying them in specific field of theme requires knowledge on the inherent characteristics of satellite data and also about the theme of application. In the present paper, a discussion on the inherent characteristics of satellite data and its utility in extracting information on the type of land units along a coastal environment has been carried out.

Keywords*:* remote sensing image (RSI), DN values, information extraction, image processing, coastal environment

1. Introduction

Remotely sensed image (RSI) and derived image databases are the fastest growing archives of spatial information that provide ample information about our earth. Tremendous amount of information hidden in these data collection play a crucial and significant role in wide range of analysis and applications. Analysis of Remote Sensing Image (RSI) is a major application domain used for various feature extraction and pattern recognition involved in natural resources assessment, hazards and environmental monitoring activities such as coastal area (Paul, 2000; Fonlupet, 2001), sea grass and mangrove ecosystem (Farid, 2002), beach morphology (Teodoro, *et.al.,* 2008) and coastal hazards (Garcin *et al*, 2008; Roemaer *et al* 2010). The process of information extraction from RSI (Yu et.al., 2000) exploit the interaction of objects on the earth with electromagnetic spectrum (ems) such as reflection, refraction and absorption, which in turn gives rise to the term spectral behavior. This spectral behavior is well exploited to identify and categorize each objects and to generate information database of any specific theme (Chen and Wang, 2004). Hence it requires an understanding on the inherent characteristics of different objects in different spectral region, attenuation or noise of signals involved due to atmospheric particles, capability of sensors and various measures of pre-processing of satellite data (Mass and Nithya, 2010) and at last types of image processing for information extraction procedures especially for coastal environment.

2. Characteristics of Remotely sensed Image (RSI)

RSI is characterized by digital values that represent the spectral reflectance of various objects as recorded by sensors on-board satellite. A digital image (of RSI) is an array of numbers depicting spatial distribution of a certain field parameters such as reflectivity of EM radiation by objects, emissivity or topographical elevation. It consists of discrete picture elements called pixels. Associated with each pixel is a number represented as digital number (DN) that depicts the average radiance of relatively small area within a scene. The range of DN values being normally 0 to 255. The size of this area effects the reproduction of details within the scene. Size of the pixel is inversely proportional to the details of a scene. For example, when the pixel size is reduced more scene detail is preserved in digital representation, which is termed as the spatial resolution of the sensor. The larger the size of the pixel, the greater the details and relatively more information about the objects of study could be obtained. That is, the spatial resolution of RSI plays a role in determining the capability of degree of information that could be obtained from a scene as well as details of an object. Similarly, the ability of sensor to record details of EM radiation of an object in narrower spectral region (bandwidth) provide ample scope to discriminate among objects as well as with in the same type of objects (Moran *et.al*., 1992).

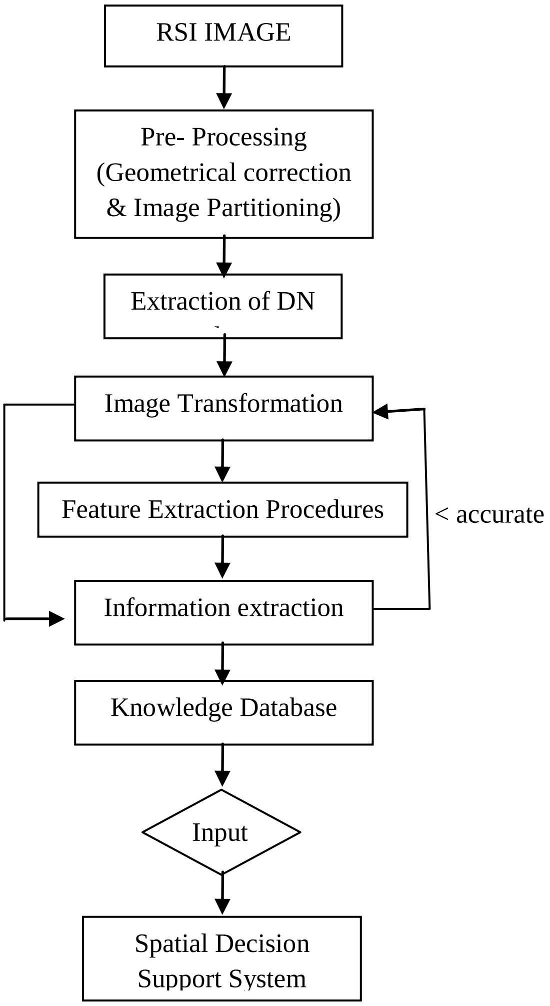

RSI IMAGE
Pre- Processing (Geometrical correction & Image Partitioning)
Extraction of DN
Image Transformation
Feature Extraction Procedures
< accurate
Information extraction
Knowledge Database
Input
Spatial Decision Support System

Figure 1. Flow chart showing information extraction from Remote Sensing Image

Thus, the spatial and spectral characteristic of RSI determine the scope of feature identification, discrimination within a specified object pattern, and details of information about objects and accuracy of information about an object.

The characteristics of RSI add complexity in processing the data in terms of size, data handling capability, feature extraction and accuracy of such information extraction. Moreover, data handling requires an understanding of data format such as Band Interleaved by Pixel (BIP), Band Interleaved by Lines (BIL) and Band Sequential Format (BSQ); the range of DN values of objects so as to segregate and identify features of area of interest as Knowledge database. Apart from these intrinsic things, effect of influence of noise on image data such as radiometric error and atmospheric attenuation apart from error due to earth rotation is to be taken care of adapting adequate measure which is termed as pre-processing the data along with generating sub-set image and tiles.

As explained above, the images of a dataset are selected according to criteria related to the application. In the preprocessing phase, feature extraction techniques are applied to these images. The mining process is a spatial data mining system prototype able to characterize spatial data using rules, compare, associate, classify and group datasets, analyze patterns and perform data mining in different levels. The extracted information may be directly integrated as “knowledge” input into any other decision making support systems. A typical information extraction procedure in RSI environment is shown in figure 1.

In the present study such a procedure is adopted to study the RSI data for extracting information of objects specifically along a coastal environment (Pais-Barbosa *et.al.*,2007). RSI is studied to extract various features such as sand, canal, river, waterbody, vegetation and saltpan based on their DN values in multispectral bands (R,G,B), textural form and pattern and position of DN values.

3. Analysis of RSI

Remote sensing image data is converted into digital number values and processed correction using Erdas Imagine software. The DN value in each band varies with the nature of objects. That is the same object will carry four DN values if the remote sensing image is a four band data image. This sort of variation in DN values in different spectral region allows the user to exploit and identify features individually for information extraction. For example waterbodies show high reflection in the first band (blue regions) and totally absorbed in the third band (red region). Any increase in DN values in the third band indicates the degree of turbidity of waterbody and presence of suspended solids in it. In this way not only the information on the identification of features is extracted but also the information on the nature of that feature as well.

In the analytical part, information extraction is the final step. The remotely sensed data is subjected to quantitative analysis to assign individual pixels to specific classes and it employs *priori* or *apriori* knowledge for categorization of pixels to some intelligent objects. Even in *apriori* approach, ground truth verification is required so as to assess the accuracy of the information derived through image processing and its reliability.

In the present study, a small tile image of a coastal area near Marakkanam town in Tamilnadu is selected and clustering technique is applied on it to analyse for extraction of possible information on the coastal features (Teodoro *et.al.,* 2009). Clustering or unsupervised approach to extract information of a coastal area would give significant reconnaissance information about the objects and could be used for further intensive studies (Zaki, . With the domain expertise available features such as canal, saltpan, sandy area, waterbody, lagoon and sea are identified from the image. The results of the analysis are discussed in the following section.

4. Results and Discussion

The resultant output as derived from applying clustering algorithm on the selected satellite data IRS-LISS III brought out information on the general setting of various land features and interaction between land and sea (Figure 2).

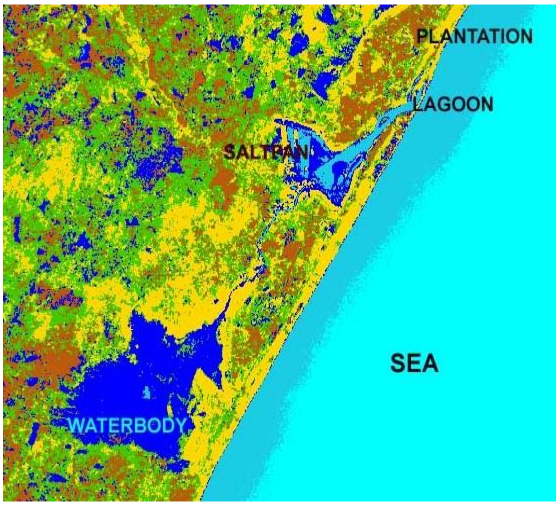


Figure 2. Resultant image of clustering

Further, to understand the significance of such information extraction, two smaller tiles have been selected and similar clustering or unsupervised procedure is applied. Such image could bring out more significant information and clearly showed more number of features. For example, lagoon area as shown at the upper part of Figure 2 (Yedayanthittu lagoon) is selected as a separate tile image and clustering algorithm is applied on it. The resultant image obtained revealed segregation of many minor features such as crop, vegetation, plantation, sand and water bodies (Figure 3). Even among the object "water body", distinguishing of turbid water along the coastline, deep sea water, and mixing of brackish water and fresh water in the lagoon could be observed.

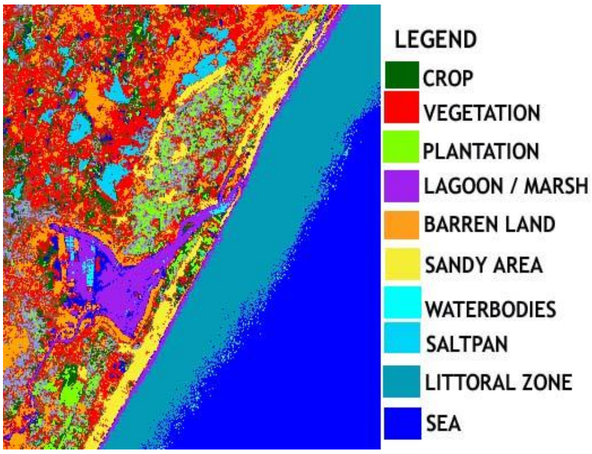


Figure 3. Resultant unsupervised image showing Lagoon ecosystem

Such variations among the water bodies as shown above is due to the varying spectral properties due to their composition (salt water and freshwater), content (sediments and soils), turbidity and mixing with vegetation. Similarly another water body, shown at the bottom part of Figure 2, a freshwater lake, is taken as the second image tile to demonstrate the significance of understanding DN values in different spectral region (Blue, Green, Red, InfraRed). A similar approach is adopted and while applying the algorithm following observations are made. Among the waterbodies, "sea", and water along the shore line called "littoral zone" and "fresh water" in the tank could be easily separated and clearly identified. There was no much confusion among pixels as the class intervals are increased. "Sand" could be easily identified and segregated as an homogenous object. This may be due to its high reflectance behaviour having high DN values compared to other objects such as "crop "and "plantation".

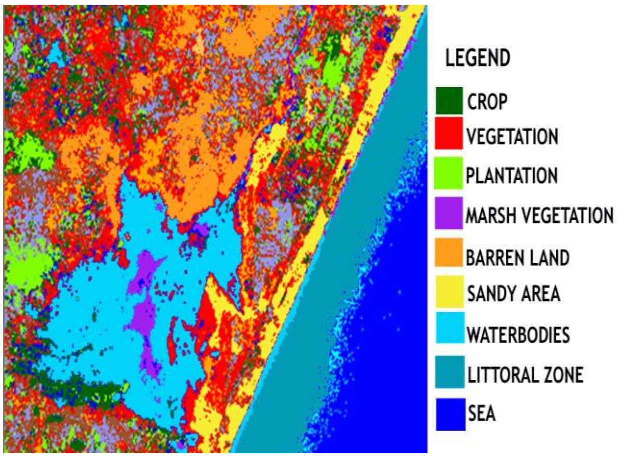


Figure 4. Resultant unsupervised image showing Freshwater ecosystem

While studying the output image, it was observed that specific feature "marsh vegetation" is seen at the middle of the fresh water. The ground truth field verification showed that small stunted growth of marshy vegetation could be observed at the middle as well at the top part of the waterbody. This may be due to the interaction of tidal water into the freshwater during monsoons and storms allowing the growth of salt tolerant floral species. Despite the use *apriori* approach, certain specific information could be brought out which otherwise may not be possible. Such a type of specific information helped to understand the degree of information extracted using clustering or unsupervised technique and at the same time enable to appreciate the interaction mechanism of nature with electromagnetic spectrum (ems), which is duly recorded in the remote sensing data. Analysis of remote sensing image in digital format and application of information extraction using unsupervised or clustering technique and observation has led to certain conclusion, which is discussed in the following section.

5. Conclusion

The analysis of RSI and observation has led to the following conclusions.

1. RSI could provide sufficient information on the land features as well coastal features.
2. Digital number (DN) values could be analysed for extracting useful information by applying appropriate processing techniques.
3. Identification and classification of objects without prior knowledge could be to some extent provide valuable information about the coastal environment.
4. Separability of different types of water bodies revealed the significance of the necessity understanding the interaction mechanism between *ems* and the objects.
5. It is possible to bring out certain specific information from clustering technique.
6. This type of study could give reconnaissance information about the coastal environment before going in for specific methods.
7. This type of approach would be more appropriate to derive baseline information about the selected study region and features along the coastal environment where predominant interaction between land and water exist.

Reference

[1] Chen, Y. and J.Z. Wang, “Image Categorization by Learning and Reasoning with Regions”, Jou. of Machine Learning Research, Vol.5, pp 913-939, 2004

[2] Farid Dahdough-Guebas, “The use of Remote sensing and GIS in the sustainable management of Tropical ecosystem”, Environment, Development and Sustainability No.4, pp 93–112. Kluwer Academic Publishers, Netherlands, 2002

[3] Fonlupt,C, “Solving the Ocean Color Problem using a Genetic Programming Approach”, Applied Soft Computing, Vol. 1:1, pp. 63–72,2001.

[4] Garcin, M., Despratts, J.F., Fontaine, M., Pedreros., Attanayake., Fernando, S., Siriwardana, C.H.E.R., De Silva, U., and Poisson, B, “Integrated approach for coastal hazards and risks in Sri Lanka”, Natural Hazards Earth Systems Science, No. 8, pp 577–586, 2008.

[5] Julea,A., M´eger,N., and Trouv´e, E, “Sequential patterns extraction in multitemporal satellite images” Proc. of 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD’06), Practical Data Mining Workshop: Applications, Experiences and Challenges, pp 94–97, 2006.

[6] Kumar, S., Ghosh, J,. and Crawford, M.M, “Best basis feature extraction algorithms for classification of hyperspectral data, IEEE Trans. Geoscience and Remote Sensing, Vol 29, No. 7, pp. 1368-1379, 2001.

[7] Moran, M.S., Jackson, R.D., Slater, P.N., and Teillet, P.M., “Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output”, Remote Sens. Env., V 41, pp 169-184, 1992.

[8] Pais-Barbosa, J., Veloso-Gomes, F., and Taveiro-Pinto, F, “Coastal features in the energetic and mesotidal west coast of Portugal, Jou. Coastal research, SI50, pp 459-463, 2007.

[9] Roemer, H., Kaiser, G., Sterr, H., and Ludwig, R, “Natural Hazards and Earth System Sciences Using remote sensing to assess tsunami-induced impacts on coastal forest ecosystems at the Andaman Sea coast of Thailand”, Nat. Hazards Earth Syst. Sci., No.10, pp 729–745, 2010.

[10] Stephan J. Maas, S.J and Nithya Rajan (2010), Normalizing and Converting Image DC Data Using Scatter Plot Matching, Remote Sens. 2010, V2, pp 1644-1661, 2010.

[11] Teodoro, A.C., Veloso-Gomes, F., and Goncalves, H., “Statistical technique for correlating TSM concentrationwith sea water reflectance using multispectral satellite data”, Jou. Of Coastal research, Vol.24, SI 3, pp 40- 49, 2008.

[12] Teodoro, A.C., Pais-Barbosa, J., Veloso-Gomes, F., and Taveiro-pinto, F., “Evaluation of beach hydromorphological behavior and classification using Image classification Techniques”, Jou. Coastal research, SI56, pp1607-1611, 2009.

[13] Yu, S., De Backer, S and Scheunders, P, “ Genetic Feature Selection Combined with Composite Fuzzy Nearest Neighbor Classifiers for High-Dimensional Remote Sensing Data”, IEEE International Conference on

Systems, Man and Cybernetics, Nashville, TN, pp. 1912–1916, IEEE Press, October 8–11, 2000.

[14] Zaki, M.J, “Spade: an efficient algorithm for mining frequent sequences”, Machine Learning, Special issue on Unsupervised Learning, 42 (1/2), pp31–60, 2001.

Authors Note

Mrs. K. Bhuvaneswari is a part-time research scholar of Mother Teresa Women’s University, Kodaikanal who is doing her Ph.D on information based approach for studying coastal environment. It is a multi-disciplinary topic involving spatial database and related techniques for environmental studies. She is a post-graduate and has completed her M.Phil in plant science. Presently she is working as PGT as well as resource person for Geosensing Information Pvt limited, Chennai.

Dr.R. Dhamotharan is a Reader in Plant Biotechnology in Presidency College Chennai, India. He has been involved in research field for the past twenty years. Presently eight candidates are doing their Ph.D under his guidance. He has published nearly twenty research papers in national and international publications.

Dr.N.Radhakrishnan, Director Geocare Research Foundation, Chennai has his doctoral degree in Spatial data techniques and its application such as remote sensing, GIS and GPS. Presently he is involved in EIA related consultancy services using RSI and GIS besides supervising research scholars doing their Ph.D in RSI domain.

Designing Flexible GUI to Increase the Acceptance Rate of Product Data Management Systems in Industry

Zeeshan Ahmed [1, 2]
[1]Vienna University of Technology Austria,
[2]University of Wuerzburg Germany.

Abstract: Product Data Management (PDM) desktop and web based systems maintain the organizational technical and managerial data to increase the quality of products by improving the processes of development, business process flows, change management, product structure management, project tracking and resource planning. Though PDM is heavily benefiting industry but PDM community is facing a very serious unresolved issue in PDM system development with flexible and user friendly graphical user interface for efficient human machine communication. PDM systems offer different services and functionalities at a time but the graphical user interfaces of most of the PDM systems are not designed in a way that a user (especially a new user) can easily learn and use them. Targeting this issue, a thorough research was conducted in field of Human Computer Interaction; resultant data provides the information about graphical user interface development using rich internet applications. The accomplished goal of this research was to support the field of PDM with a proposition of a conceptual model for the implementation of a flexible web based graphical user interface. The proposed conceptual model was successfully designed into implementation model and a resultant prototype putting values to the field is now available. Describing the proposition in detail the main concept, implementation designs and developed prototype is also discussed in this paper. Moreover in the end, prototype is compared with respective functions of existing PDM systems .i.e., Windchill and CIM to evaluate its effectiveness against targeted challenge.

Keywords: Human Computer Interaction; Flexible Graphical User Interface, Product Data Management, PDM System, Prototype Development, Rich Internet Applications

1. Introduction

Companies consist of different departments like management, marketing, accounts, production, quality and engineering etc. Every department has its own rules, regulations, data and staff. There is no doubt every department is important and expected to play a vital role in the progress of industrial enterprises but most important of all is the engineering or technical department, which is more responsible for the main product's development and production from all other departments. To successfully run the technical department hardware and software are deployed, processes are initiated and implemented, required number of technical staff is hired to produce the product under the implemented process using the provided resources. Problems initiate and start growing as a company grows due to the rapid increase in data with the lack of required in time information and project and resource management. As a result the company can face unnecessary additional increase in costs, delays in product completion, loss in quality and waste of time [7].

In the past, there were no such systems available to store, track and manage all the related product data. This doesn't mean that there was no system for data management; there were some systems to store the information about product, personnel involved in organization and financial details but there was no such comprehensive system to manage technical data. To cope with the problem of organizational technical data management a new system category was introduced i.e., Product Data Management (PDM). PDM is a digital way of maintaining engineering data of technical departments within organizations to improve the quality of products and processes. PDM products mainly manage information about design and manufacturing of products including technical operations and running projects.

Till now everything sounds perfect, but the problems initiate and start growing as company grows. These problems can happen because of lack in control over engineering processes, rapidly increasing data, lack of presence, lack of coordination among team members (staff), unclear product configurations, loss of experienced staff, conflicts between the central Information Systems (IS) organization, lack of suitable formal communications between departments, bureaucratic and complex engineering change control systems and lack of project and resource management. As the result company can face unnecessary additional increase in cost, delays in product completion, loss in quality and waste of time.

Successful PDM System Deployment in an organization (especially large one) is quite difficult because it is time consuming, expensive and most of the staff (belonging to corporate management, top level management, engineering management and other engineering and IT professionals) do not give importance to PDM System and without these person's support it is quite difficult to implement it. Moreover people don't want to involve in low level technical and business issues, don't want to spend money, look for fast pay back projects, don't have extra time, too much inertia in this company, lack of trust of users on management, job insecurity and incapable of handling PDM systems.

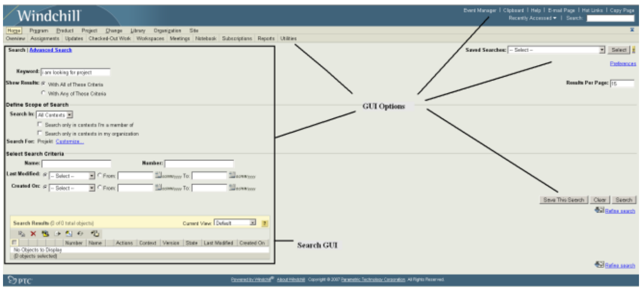
Some of the main reasons of lack in acknowledgement of PDM Systems in international market are some problematic issues and if these are resolved then it will be a great contribution to PDM System development, usage and marketing. The graphical user interfaces of most of PDM Systems are user unfriendly, nonflexible and slow (especially if the system is web based). In case, if PDM System is a client based application then the issue of platform independency is also there because in the new business models it is nearly impossible to mandate that all the potential users choose the same platform or the same operation system. Moreover PDM Systems normally deal with heavy amount of data but in most of the cases it is quite difficult to access or search needed information by using intelligent search mechanism. Without going into the details of all PDM System problems and residing within the scope of this research, focusing only one of the all current industrial unresolved issues in PDM System development i.e., unfriendly graphical user interface.

If a product is very productive and with lots of beneficial functionalities but if it is not easily usable then in most of the cases it becomes a flop in industry. Designing and implementing an intelligent and user friendly HMI for any kind of software or hardware application is always a challenging task for the designers and developers because it is very difficult to understand the psychology of the user, nature of the work and what best suits the environment. Normally PDM Systems offer many different services and functionalities at a time but the graphical user interfaces of most of the PDM Systems are not designed in a way that a user (especially a new user) can easily learn and use them. Most of the web GUI of PDM applications are with massive control implementation at user end, providing several options at a time which might not be in need of every user. Moreover the GUI of PDM applications are not flexible enough that a user can change the default orientation and placement of controls according to his need and choice.

The goal of this research is to support PDM with a web based platform independent PDM approach capable of providing flexible graphical user interface. This research is about to propose a new flexible web based GUI for multiple roles based client PDM systems capable of providing faster and better access of the system, options to change the default orientation of provided GUI controls, add or delete provided options, even a user can redesign a new graphical look by changing the default GUI according to his need and wish. Continue the discussion with the identification and detailed presentation of Problem Definitions in section 2. Then in section 3 and 4 Human Computer Interaction and Rich Internet Application are presented, as the part of state of the art. Newly proposed Approach towards some of the PDM System problems is presented in section 5 of this research paper. Narrowing the scope of this research paper and focusing only on the proposing of new flexible web based graphical user interface for PDM Systems, a new approach, its concept, implementation designs and developed Prototype is presented in sections 6, 7, 8 and 9 of this research paper. Later in section 10 of this research paper, the resultant prototype is compared with some existing PDM Systems, to evaluate its effectiveness. In the end, some existing limitations in developed prototype are presented in section 11 and discussion is concluded in section 12 of this research paper.

2. Problem Definition

PDM Systems offer different services and functionalities at a time for many types of roles/users like managers, designers, engineers etc. but the graphical user interface of most of the PDM Systems are not designed in a way that a user (especially a new user) can easily learn, use and adopt [6]. Even at times for the old users it becomes a massively complicated GUI with several options from which many of them are not even in use all the time. Moreover the GUI of PDM applications are not flexible enough that a user can change the default orientation of controls by redesigning the default GUI according to his need and choice, and can save it so that it can be reused later on.


Windchill
Event Manager | Clipboard | Help | E-mail Page | Hot Links | Copy Page
Recently Accessed | Search:
Home Program Product Project Change Library Organization Site
Overview Assignments Updates Checked-Out Work Workspaces Meetings Notebook Subscriptions Reports Utilities
Search | Advanced Search
Saved Searches: -- Select -- Select
Preferences
Results Per Page: 15
Define Scope of Search
Search In: All Contexts
Search only in contexts I'm a member of
Search only in contexts in my organization
Search For: Projekt Customize...
Select Search Criteria
Name:
Number:
Last Modified: -- Select -- From: To:
Created On: -- Select -- From: To:
Save This Search
Clear
Search
Refine search
Search Results (0 of 0 total objects)
Current View: Default
Number Name Actions Context Version State Last Modified Created On
No Objects to Display
(0 objects selected)
GUI Options
Search GUI
PTC

Keyword: i am looking for project
Show Results: ● With All of These Criteria ○ With Any of These Criteria

Figure 1. Windchill (Marked) Graphical User Interface

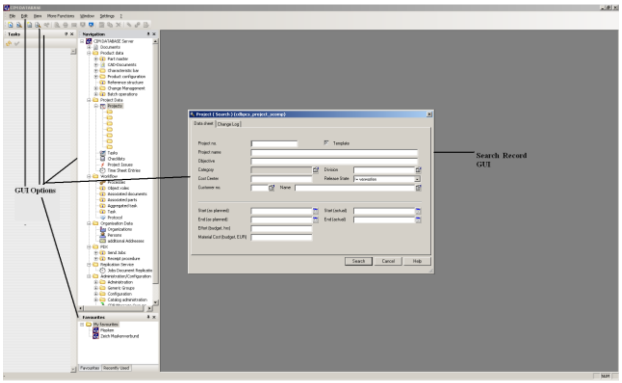


Figure 2. CIM Data Base (Marked) Graphical User Interface

PDM Systems are especially designed for the role based client users (multiple users playing different roles with different rights in the same organizations). So the probability of predicting that every user does not need all the options of the system all the time is very high. Moreover we can also say that the massive availability of all the controls to all the users all the time will also reduce the speed and efficiency in the work of the users because if a user will only be provided with some limited options with respect to his nature of job, rights and responsibilities then it will be much faster and more convenient for him to use and perform in the system. Moreover if the user is also provided with the flexibility of redesigning GUI by changing the default orientation of the provided controls (by adding or deleting) in default GUI of the PDM system according to his ease and the need then it will also be a useful contribution at the user's end. Currently available PDM Systems e.g. Windchill [1] as shown in Figure 1 and CIM Database [2] as shown in Figure 2 are excellent engineering data management systems but with not user friendly and flexible graphical user interface with structured input mechanism and massive input/output controls.

2.1 Example of Unfriendly GUI; Windchill

Windchill is a web enabled product data management application. It provide services in content and process management of organizational, technical and managerial data as shown in figure 1. Windchill is capable of decreasing product development time through efficient collaboration, reducing errors by automating processes and driving conformity to corporate standards, reducing scrap and rework by automatically sharing product data with downstream manufacturing systems and engineers, increasing efficiency by enabling engineers to quickly find and manage multiple forms of digital product content, eliminating mistakes associated with duplicate data, incomplete data, or manual data and technology risk by reducing the number of systems and databases to maintain and administer.

Windchill is heavily benefiting the industry by providing such an excellent engineering data management system at a time it is providing difficulties to its clients in using the system by providing a massive graphical user interface with more than 30 main GUI control options e.g. Program, Product, Project, Change, Library and Organization etc. and more than hundreds of sub main GUI control options at a time. Because of this massive provision of GUI controls it is difficult for the new user of Windchill to learn and adopt it quickly, and for the old users it is sometimes complex to search required main or sub main GUI options. Furthermore the GUI of the Windchill is not flexible so a user can not even make a little alteration according to his need and wish. Apart from the above discussion the currently available version of Windchill 9 is also very slow and this is another factor to be looked at and improved for better use.

2.2 Example of Unfriendly GUI; Windchill

CIM Database is a Client Server PDM with native client and an optional Web client to manage engineering data and support product creation and process implementation as shown in figure 2. CIM Database is a secure data management system for a range of centralized functions like selectable search, CAD systems, product and organizational data management and electronic data interchange (EDI). Likewise Windchill the CIM Database is also heavily benefiting the industry by managing engineering data and at the same time providing massive graphical user interface with more than 50 main GUI control options e.g. Product Data, Project Data, Workflow, Organization Data, PDX, Replication Services, Administration/Configurations and More Functions etc. and more than hundreds of sub main GUI control options at a time. But unlike Windchill the GUI of the desktop based client application of CIM Database is more user friendly and flexible.

As it is a desktop based client GUI, it provides some options in GUI control alignment and orientation e.g. user can change the placements of provided control trays, faster and with a better access to the controls etc. But apart from these advantages there is a big disadvantage of desktop based client GUI that it is not available using world wide web. The desktop based client must need to be installed before using this PDM System. Moreover using CIM Database a user is also restricted to perform only one task (while making search) at a time because CIM Database desktop based client search module is based on Single interface Data Input (SDI) concept. To overcome these deficiencies CIM has also launched a web based client GUI but with the almost same limitations of GUI earlier mentioned in Windchill's GUI discussion e.g. user unfriendliness and nonflexibility etc.

As discussed earlier and shown in figure 1 and 2 the GUI of Windchill and CIM Database is more or less same like traditional database applications consisting of several options like data manipulation forms to enter or edit or delete data, search forms to find needed information, print information, use of CAD for making designs etc. Moreover CIM Database and Windchill consists of massive (providing several options to each user which might not be needed every time but still they are there), nonflexible (GUI is not flexible enough; a user cannot change the orientation of controls according to his need and choice) and user unfriendly GUI (massive controls and non flexible GUI these are not much user friendly and it is quite difficult for a new user to adopt to them). Because of these deficiencies in the GUIs of the existing PDM Systems, a flexible web based GUI for the multiple roles based client PDM Systems is need to be proposed which should provide faster and better access of the System to the users by providing options to the users to change the default orientation of provided GUI controls according to the need and wish, better access to the controls, user's own choice look and feel which user can design, redesign, save, use and later can alter as well.

3. Human Computer Interaction

Targeting the challenge of proposition of designing a flexible web based graphical user interface development; I have chosen the field of Human Computer Interaction (HCI) to have complete understanding of graphical user interface design and development. HCI is the study of design, evaluation and implementation of interactive computing systems for human use [3]. Designing High quality HCI design is difficult to implement because of many reasons .i.e., market pressure of less time development, rapid functionality addition during development, excessive several iterations, competitive general purpose software and human behavior analysis.

Designing human computer interaction interface is an important and a complex task, but it could be simplified by decomposing task into subcomponents and maintaining relationships among those subcomponents. Task decomposition is a structured approach, applicable in both Software Engineering and Human Computer Interaction (HCI) fields depending on specific processes and design artifacts. Using design artifacts applications could be made for analysis and design by making the hand draw sketches to provide high level of logical design based on user requirements, usage scenarios and essential use cases. To design hand draw sketches there are some strategies to be followed .i.e., planning, sequential work flow, and levels of details.

3.1. HCI Design Principles

While evaluating or designing a user interface, it is important to keep in mind the HCI design principles. There are four major HCI design principles .i.e., Cooperation, Experimentation, Contextualization, Iteration and Empirical Measurement [4].

1. Cooperation plays a vital role in software project development. The most important and primitive principle of design process is the cooperation between both developers and the end users. Because in the design process with respect to the participatory design point of view there exists an uncommon principle .i.e., presenting the same issues with completely different perspectives and dimensions.
2. Generally experimentation is performed in the middle of recently acquired possibilities and the currently existing conditions. To assure that the present conditions are in conjunction with new ideas and supported by two primitive principles .i.e., concretization and contextualization of design, Principles are in associated with the above mentioned visions performing experiments with visions and hand on experience.
3. Design hooks its initial point with a particular configuration in which new computer based applications put into practice. Participatory design emphasizes on situations based on the implementation of iterative designs. The design composition of use is tied up with numerous social and technical issues. Generally participatory design of the development will specifically includes different kinds of participants i.e. Users, Managers and the design developers.
4. In design process, hang on to some issues which are not yet revealed, which are visioning the future product from design point of view and the construction of work from use point of view. But participatory design puts a controversial statement in accomplishing the same by making use of artifacts i.e. Prototype. Designers with cooperation will make use of the artifacts as a source for delegation of work. Participatory design also ends up with a controversial statement for trivial division of work in the process of development, which pleads overlap among the members of analysis, design and realization groups.
5. Empirical measurement is about to test the interface in early stages with the involvement of real users who come in contact with the interface on an everyday basis. Keep in mind that results may be altered if the performance level of the user is not an accurate depiction of the real human computer interaction. Furthermore its also about to establish quantitative usability specifics such as: the number of users performing the task, the time to complete the task and the number of errors made during the task.

3.2. Design Patterns

Like software engineering design patterns there are some graphical user interface design patterns i.e., Window Per Task, Direct Manipulation, Conversational Text, Selection, Form, Limited Selection Size, Ephemeral Feedback, Disabled Irrelevant Things, Supplementary Window and Step-by-Step Instructions. These patterns help designers in analyzing already designed graphical interfaces and designing a user friendly and required on demand graphical machine interface [3] e.g.

- Window per task helps in organizing the complete graphical user interface into different screens by providing the information about tasks per window screen.
- Direct Manipulation is a user machine interaction style where user interacts with system by directly using provided options.
- Conversational Text provides textual input information of designed interface's commands.
- Selection describes interaction style to choose options from provided list of options.
- Form describes discrete structures on screen.
- Limited Selection Size structures set of selections.
- Ephemeral Feedback provides the information about the natural flow of the interface.
- Disable Irrelevant Things guide in identifying and removing irrelevant interface elements.
- Supplementary Window provides information about supplementary windows.
- Step by Step Instructions help designer in sequencing set of actions

3.3. HCI Design Guidelines

A successful design interface can be implementable using the following guidelines .i.e.,

- Design mock ups should be implemented.
- Design should be presentable according to the need of the user.
- Criteria / principles should be applied to the design.
- Prepared according to the project proposal based on specified functional requirements.
- Should be evaluated with respect to the number of features asked to develop.
- Assessed by testing especially in work load conditions.
- Use case modeling should be used with the identification of user interface elements
- Should be flexible enough to adopt rapid prototype changes and modifications.
- Should be based on consistent sequences of actions required in similar situations.
- Should be based on identical terminologies used in prompts, menus, and help screens.
- Should be based on consistent color, layout, capitalization, fonts, and so on should be employed throughout.
- In case of massive GUI based many components, HCI should enable frequent users to use shortcuts o increase the pace of interaction with the use of abbreviations, special keys, hidden commands and macros.
- Provide informative feedback for every user action.
- Should categorized sequences of actions into groups.
- Should offer error prevention and simple error handling.
- Should provide permit easy and reversal of actions.
- Should reduce short term memory load
- The GUI should provide an obvious, intuitive, and consistent interface to the simulation system.

- The GUI should provide an efficient means for reusing component models.
- The GUI should provide different graphical layouts for different types of simulation applications.

4. Rich Internet Application

The term "Rich Internet Application" was introduced in a white paper of March 2002 by Macromedia. Rich Internet Applications (RIA) are web applications with features and functionalities of traditional desktop applications as well as web applications. Traditional web applications center all activities around client server architecture with a thin client where as RIA typically transfer the processing necessary for the user interface to the web client but keeps the bulk of the data (i.e., maintaining the state of the program) back on the application server.

RIA shares one characteristic with other web development technologies, an intermediate layer of code often called a Client Engine, between the user and the server. This client engine is usually downloaded as part of the instantiation of the application, and may be supplemented by further code downloads as use of the application progresses. The client engine acts as an extension of the browser, and usually takes over responsibility for rendering the application's user interface and for server communication. Using Client Engine RIA becomes richer, more responser, balanced, asynchronous and efficient.

- *Richness*: User interface behaviors are not obtainable using only HTML widgets available to standard browser based Web applications. This richer functionality may include anything that can be implemented in the technology being used on the client side, including drag and drop, using a slider to change data, calculations performed only by the client and not needing to be sent back to the server.
- *Responsively*: The interface behaviors are typically much more responsive than those of a standard Web browser that must always interact with a remote server. The most sophisticated examples of RIA is that it exhibits a look and feel of a desktop environment level. Using a client engine can also produce other performance benefits.
- *Balanced*: The demand for client and server computing resources is better balanced, so that the Web server needs not to be the working horse like in traditional Web application. This frees server resources and allows the same server hardware to handle more client sessions concurrently.
- *Asynchronous*: The client engine can interact with the server without waiting for the user to perform an interface action such as clicking on a button or link. This allows the user to view and interact with the page asynchronously from the client engine's communication with the server. This option allows RIA designers to move data between the client and the server without making the user wait. Perhaps the most common application of this is pre-fetching data, in which an application anticipates a future need for specific data and downloads it to the client before the user requests it, thereby speeding up a subsequent response. Google Maps use this technique to load adjacent map segments to the client before the user scrolls them into view.
- *Efficiency*: The network traffic may also be significantly reduced because an application-specific client engine can be more intelligent than a standard Web browser while deciding which data needs to be exchanged with servers. This can speed up the individual requests or responses because less data is being transferred for each interaction, and overall network load is reduced. However, over-use of asynchronous calls and pre-fetching techniques can neutralize or even reverse this potential benefit because the code cannot anticipate exactly what every user will do next, it is common for such techniques to download extra data, not all of which is actually needed, to many or all clients.

4.1. RIA Technologies

There are several RIA technologies available i.e., FLEX (Adobe), AJAX, OpenLaszlo and Silverlight (Microsoft).

- Flex is a highly productive, free open source framework for building and maintaining expressive web applications that deploy consistently on all major browsers, desktops, and operating systems. While Flex applications can be built using only the free Flex SDK, developers can use Adobe Flex Builder™ 3 software to dramatically accelerate development. Adobe Flex is a collection of technologies released by Adobe Systems for the development and deployment of cross platform rich Internet applications based on the proprietary Adobe Flash platform.
- AJAX is a free framework for quickly creating efficient and interactive Web applications that work across all popular browsers. AJAX stands for Asynchronous JavaScript and XML. AJAX is a type of programming which became popular in 2005 by Google. It is not a new programming language, but a new way to use existing standards. Its primary characteristic is the increased responsiveness and interactivity of web pages achieved by exchanging small amounts of data with the server "behind the scenes" so that entire web pages do not have to be reloaded each time, there is a need to fetch data from the server.
- OpenLaszlo is an open source platform for the development and delivery of rich Internet applications. It is released under the Open Source Initiative-certified Common Public License. Laszlo applications can be deployed as traditional Java servlets, which are compiled and returned to the browser dynamically. This method requires that the web server be running the OpenLaszlo server. OpenLaszlo was originally called the Laszlo Presentation Server (LPS).
- Microsoft Silverlight is a web browser plug-in that provides support for rich internet applications such as animation, vector graphics and audio-video playback. Silverlight provides a retained mode graphics system,

similar to WPF and integrates multimedia, graphics, animations and interactivity into a single runtime. It is being designed to work in concert with XAML and is scriptable with JavaScript. XAML can be used for marking up the vector graphics and animations. Textual content created with Silverlight is more searchable and indexable than that created with Flash as it is not compiled, but represented as text (XAML). Silverlight can also be used to create Windows Sidebar gadgets.

Table 1. Comparison between RIA Technologies

Content	Flex	Silverlight
IDE GUI	Yes	Yes
Project User Interface declarations	XML based (MXML)	XML based (MXML)
Cross-platform	Yes	Windows Only
Server side integration	object based, AMF	object based, AMF
Worldwide usage	Best	Poor
Loading time / Boot	Fast	Good
3D	HW supported	HW supported
Components & Tools	Better	Good
Component integration with OS	Good	Bad

Based on the earlier discussed RIA based information and to conclude with one final technology for own flexible web based graphical user interface development, a comparison is performed between two most beneficial technologies of all i.e., Flex and Silverlight. As the result of comparison, on the basis of above presented results in table 1, Flex is chosen for the own flexible web based graphical user interface development for PDM Systems because Flex has the biggest advantage of being used for cross platform (operating system independent), having fast loading time and with provision of better tools and components.

5. Proposed Approach

Focusing on the need of an approach as the solution towards the problems of implementing a flexible graphical user interface for PDM System development, I have chosen and thoroughly investigated the field Human Computer Interaction, based on the resultant information of conduced research; a new approach has been proposed. As shown in Figure 3, the proposed approach consists of four different modules i.e. Flexible GUI, NLP Search, Data Manager and Data Representer.

Proposed approach is mainly for the development of a PDM system capable of providing a flexible web based graphical user interface, identifying user's structured and unstructured natural language based requests, processing natural language based user's requests to extract results from attached repositories [5], manage data in database management system and represent system outputted information as the result of user input in user's understandable format. In this research paper without going into the details of other three modules of proposed approach, will only discuss the module i.e. Flexible GUI.

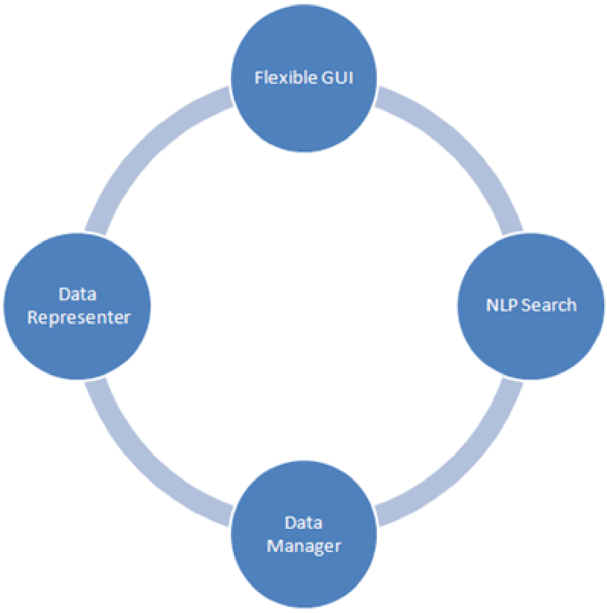


Figure 3. Conceptual Model of Proposed Approach

6. Proposed Flexible GUI

As shown in Figure 4, different kinds of users i.e., Businessman, Project Manager, Engineer and staff member etc. are need to interact with PDM System at a time. The major interests for a Businessman could be regarding the performance and quality of running projects. Project manager's job is to plan function including defining the project objective and developing a plan to accomplish the objective, organizing function involves identifying and securing necessary resources, determining tasks that must be completed, assigning the tasks, delegating authority, and motivating team members to work together on the project and manage running projects. Engineer is there to design product using CAD whereas other staff members could be involved in different tasks e.g. organization's personal and project data entrance and management etc. These different kinds of users have different kinds of psychologies to approach and use one PDM System. As PDM Systems consists of different options and these are designed and implemented for different kinds of users. The point to think is how a PDM system can provide a user system interaction mechanism which can satisfy all kinds of users because it is quite difficult to provide one graphical user interface which can satisfy all users by providing their needed components without creating a mess of options at GUI. Keeping this need in mind, proposed a new approach i.e. Flexible GUI, for PDM System development.

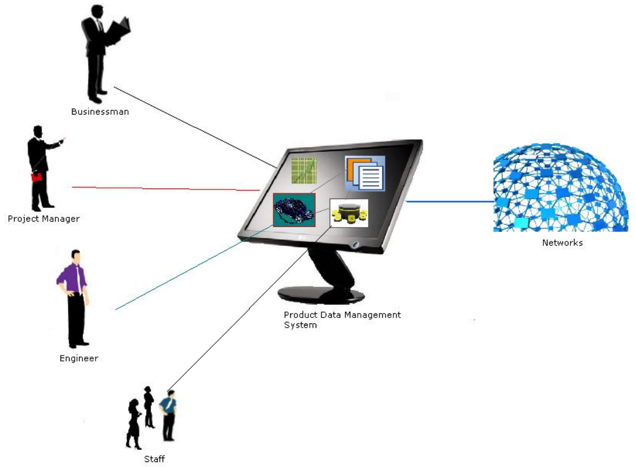


Figure 4. PDMs Multi Role based Users System Interactions

To manage the complex control flows necessary for GUIs designed for PDM Systems with flexible interfaces, a new way of Flexible GUI implementation is presented. Flexible GUI is a type of user interface that allows user to interact with the program in more ways than typing such as computers hand held devices and office equipment with images rather than text commands. Flexible GUI uses a combination of technologies and devices to provide a platform independent user interface which any user can interact with for the respective tasks. The design of Flexible GUI is based on three properties i.e. User friendliness, Model reusability and Application extensibility. User friendliness provides obvious, intuitive, and consistent interface, Model reusability provides an efficient means for reusing developed component and Application extensibility provides different graphical layouts for different types. Furthermore Flexible GUI's structure is flexible enough to accommodate graphical layout for different kind of user of different applications.

Flexible GUI is mainly a friendly web based graphical user interface proposed for product data management systems for better the human computer interaction. The main idea behind the proposition of a new web based graphical user interface is to improve user system communication by providing several options helping user by letting him change the default orientation of the GUI by changing the placements of provided controls, insertion of needed and deletion of unnecessary controls and redesigning completely new look and feel of the GUI, which is not at the moment possible in almost every PDM System.

Targeting the problem of a user friendly graphical user interface, the proposed flexible graphical user interface is designed keeping the need of provision of different services and functionalities at a time for many types of roles/user in mind. The proposed Flexible GUI for PDM applications is flexible enough that a user can change the default orientation of controls by redesigning the default GUI according to his need and choice, and can save it so that it can be reused later on because the GUI of most of PDM Systems is massively complicated with several different options at the same time to all the users from which many of them are not even in use of all the user at all the times. Furthermore the proposed Flexible GUI is especially designed for the multiple roles based clients providing faster and better access of the System.

7. Flexible GUI; Conceptual Designs

Following information obtained as the result of conducted research in the field of Human Computer Interaction, a mockup (draft physical sketch) of proposed Flexible GUI is designed for a prototype development of proposed approach. The mockup is presentable according to the need of the user, designed with respect to the criteria and principles followed by the system and flexible enough to adopt rapid prototype changes and modifications. The mockup is based on an interactive design displaying required quantitative material including images, windows and tools etc.

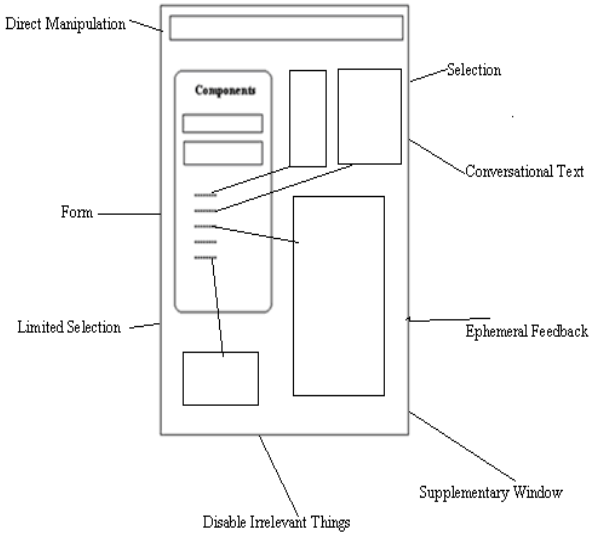


Figure 5. Mockup of Proposed Flexible GUI

The mockup of proposed Flexible GUI is needed to be implemented in the form of three different web pages using RIA technologies i.e. Components Page. The Component page is a prototype form of a flexible web based graphical user interface, consisting of a control container, giving an idea for placing all the components based options involved in the Product Data Management operations in a user's desired way by adding or deleting provided options. Further Component page also allows the user to redesign web based GUI with respect to its own choice by changing the GUI orientation by altering the GUI Component placements, changing the size of GUI control components (e.g. list boxes, mouse hover/click, drag drop, drop down list boxes, list boxes etc.) and changing the used color scheme, font, background etc. of in use GUI. The mockup of Component Page, as shown in Figure 5 is based on eight design patterns i.e. Direct Manipulation, Conversational Text, Selection, Form, Limited Selection, Ephemeral Feedback, Disable Irrelevant Things, Supplementary. Project implementation designs are created using these mockups for the prototype implementation of Flexible GUI using RIA technologies.

8. Flexible GUI; Implementation Designs

8.1. Design Methodology

Following three the classical tier application model, I have designed implementation methodology for the development of proposed prototype, as shown in Figure 6. The current version of proposed approach will be implanted with the use of Java (servlets and JSP) to handle user input, manage and retrieve data from the database. Tomcat is used as the main web server and middleware of the program. Users can access the web pages with the given URL and then can build graphical user interface or search the data after successful identity authorization. The data communication between three tiers is managed by Action Message Format (AMF) using the Simple Object Access Protocol (SOAP). AMF based client requests are delivered to the web server using Remote Procedure Call (RPC). The use of RPC allows presentation tier to directly access methods and classes at the server side. When data is request from user then a remote call is made from the user interface in the remote services' (via the server side includes) class members and the result is sent as an object of a Java class.

A web browser is mainly needed to access the developed application with a user of a specified universal resource link (URL). User will send a request to the web server through Hypertext Transfer Protocol (HTTP), the web server will pass the request to the application components. These application components are implemented using servlet/JSP, designed to handle user request coming from web server with the use of java remote classes. Then used servlets or JSP classe talks to the database server, perform the data transactions and send the response to the client. To increase flexibility of graphical user interface at client end, the development of front end is performed using Flex Flex (Builder 3 IDE), Relational database is designed and implemented using MySQL 5.

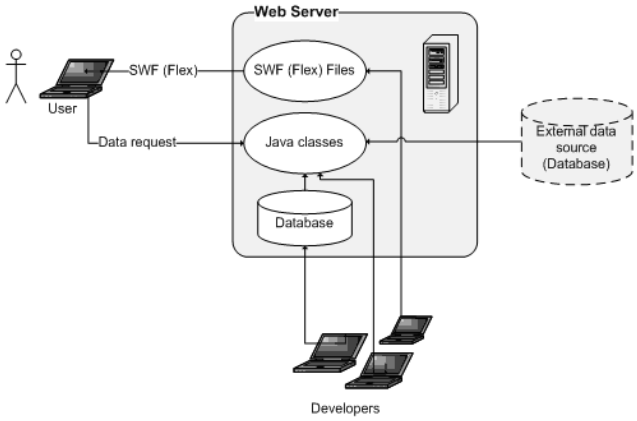


Figure 6. Implementation Design

8.2. System Sequence

As shown in figure 7, the Sequence design of the Flexible GUI consists of three components .i.e., Default GUI, Flexible GUI and Store GUI Setting.

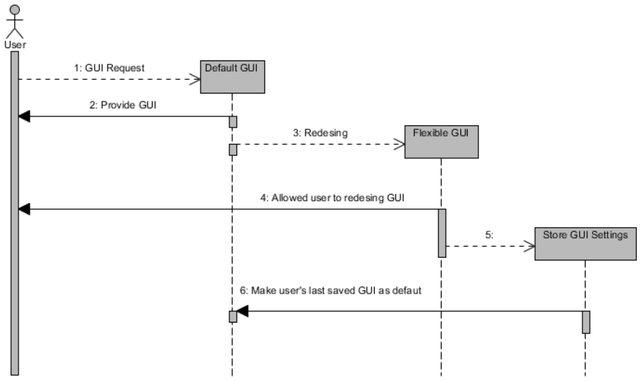


Figure 7. System Sequence Design

The job of Default GUI is to first identify user and then provide default system graphical user interface to the user, and incase a new graphical user interface is already designed and stored by user, then to provide his previously stored graphical user interface. Furthermore it also allows user to redesign a new graphical user interface with respect to this choice using providing components.

9. Flexibel GUI: Prototype

Following the constructed mockup, implementation designs, meeting the design requirements for a proposed Flexible GUI and residing with in the limited scope of this research's development, a prototype version of proposed approach is developed with the use of RIA technologies. This prototype version is Web application is capable of providing flexible graphical user interface with several different options (for multi role based clients) for Product Data Management Systems. The flexible web based graphical user interface is developed following designed mockup and divided into two sections as shown in Figures 8 and 9 i.e., Default Graphical Interface and User Graphical Interface.

Figure 8. Prototype; Default Graphical Interface

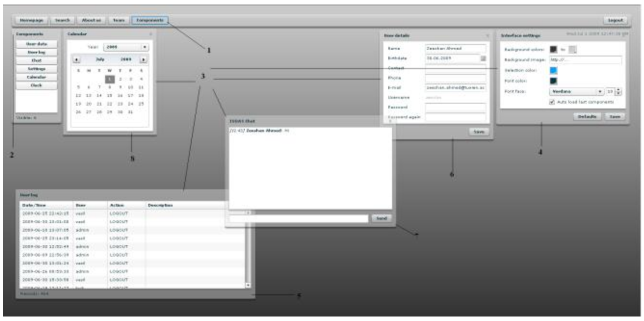


Figure 9. Prototype; User's Personal Graphical Interface

The default graphical interface is the graphical interface with some limited and basic options which can be accessed by the user and the guest. But the User graphical interface can only be accessed by the user after logging into the application with authenticated user name and password. User graphical interface is the actual interface presenting the prototype definition of proposed idea about Flexible GUI in conceptual design, as shown in figure 9 and described in table 2.

Table 2. Graphical Interface Page

No	Option	Description
1	Main Component Link	To enable the Graphical Interface
2	Components Tray	Providing all GUI options to the user to click and use
3	Components	All currently available components for PDM operations and GUI manipulations
4	Interface Setting	This components if used to change the outlook of the graphical interface by changing the color schemes and adding or removing image to the main interface
5	User Log	This components provides the detail of all the operations performed during the use of graphical interface, but this components if only visible to the user with administrative rights
6	User Details	This components provides the options to enter and alter user details
7	Chat	This component is providing option for in house chat to the login users to improve in house communications
8	Calendar	This is the simple calendar to enable user with date.

This implemented prototype version is capable of

- Providing standard graphical interface designed by system.
- Providing flexible graphical user interface, so the user can redesign and reconfigure the interface itself to accommodate specific needs by Mouse Click and Drag Drop options.
- Providing several options to the user for GUI designing like user can change the look and feel by changing background colors, font and images, adding, deleting and altering components.
- Providing option to every user to save his own deigned GUI, so that the next time if the user comes online then he will be provided his own designed GUI rather than the default one. However he will still have the option to redesign or alter or restore the default GUI.

10. GUI Comparison

10.1. Prototype with CIM Database

The presented results in table 3 of performed comparison between the GUIs of prototype and CIM database demonstrates the contributions of prototype's GUI towards the PDM systems with respect to the scope, goal and earlier discussed defined problematic GUI based issues. The GUI of the client based desktop CIM database is quick and efficient in providing fast and easy access to provide the controls but at the same time it is not platform independent, it is not flexible enough so then a user can change the orientation of controls and can redesign GUI according to his choice and will but on the other hand the implemented prototype version of prototype is capable of providing these missing features in quick and efficient way.

Table 3. GUI Comparison between Prototype and CIM Database

No	Jobs	CDB	Prototype
1	Web based graphical user interface	No	Yes
2	Platform independent graphical user interface	No	Yes
3	Default GUI designed by system	Yes	Yes
4	User based Flexible Graphical Interface	No	Yes
5	The orientation of controls at GUI can be changed.	Yes	Yes
6	Reoriented controls of GUI can be saved and reused	No	Yes
7	Outlook of graphical user interface can be changed or newly designed.	No	Yes
8	Newly redesigned user based graphical interface can be saved and altered again.	No	Yes
9	Quick and efficient control's movement and data presentation.	Yes	Yes

Prototype's GUI is platform independent, flexible enough so a user can redesign the default GUI by adding or deleting provided controls, changing the placements of in use controls, modifying the outlook of GUI according to his own choice and will and saving new redefined GUI for later reuse.

10.2. Prototype with Windchill

The presented results in table 4 of performed comparison between the GUIs of Prototype and Windchill database demonstrates the contributions of Prototype GUI towards the PDM systems with respect to the scope, goal and earlier discussed defined problematic GUI based issues. The GUI of the Windchill is web based platform independent application and with all needed options for engineering data management but at the same time if compared with Prototype's GUI then its GUI is slow, not flexible that a user can not change the orientation of controls and cannot redesign GUI according to his choice. Moreover in case of Windchill user is restricted to only use the default GUI with provided massive controls even when he is not in need of many of them. But in case of Prototype, the provided GUI is platform independent and flexible so that a user can redesign the default GUI by adding or deleting provided controls, changing the placements of in use controls, modifying the outlook of GUI according to his own choice and will and saving new redefined GUI for later reuse.

Table 4. GUI Comparison between Prototype and Windchill

No	Jobs	CDB	Prototype
1	Web based graphical user interface	Yes	Yes
2	Platform independent graphical user interface	Yes	Yes
3	Default GUI designed by system	Yes	Yes

4	User based Flexible Graphical Interface	No	Yes
5	The orientation of controls at GUI can be changed.	No	Yes
6	Reoriented controls of GUI can be saved and reused	No	Yes
7	Outlook of graphical user interface can be changed or newly designed.	No	Yes
8	Newly redesigned user based graphical interface can be saved and altered again.	No	Yes
9	Quick and efficient control's movement and data presentation.	No	Yes

11. Limitations

The initial plan was to implement maximum possible PDM functionalities during the development of Flexible GUI of proposed approach but due the time limitations and limited scope of this research, development was restricted to the implementation of some of the functionalities putting some values but giving good idea that how can a complete Flexible GUI be implemented for a PDM System with all components and functionalities needed for a complete PDM System.

12. Conclusions

Targeting the challenge of proposition of web based flexible graphical user interface development; a thorough research has been conducted in Human Computer Interaction and RIA Technologies. Taking help from observed information from conducted research in respective field and using person research and development experience, I have proposed an approach. I have designed conceptual and implementation designs of proposed approach and implemented it using some software tools and technologies of present time i.e. Flex, Java, Antlr, MySQL, and presented developed prototype solutions.

In the end concluding the research and development efforts, we can say that proposed approach can put some values in enhancing PDM System development process by highlighting some existing challenges in PDM System development and proposing a new idea (along with conceptual and implementation designs) for flexible graphical user interface development to professional PDM System developing organization e.g. Windchill, CIM etc. The inclusive implementation of this proposed idea in PDM System development can put some values in increasing the market values of PDM Systems by increasing its acceptability in industry by improving its use amongst managerial, technical and office staff, because I strongly believe that if a product is very productive and with lots of beneficial functionalities (like PDM Systems) but not easily adoptable by its users then in most of the cases it becomes a failure in industry.

13. Acknowledgments

I am thankful to University of Wuerzburg Germany and Vienna University of Technology Austria for giving me the opportunity to keep working on this research project. I am thankful to Prof. Dr. Detlef Gerhard for his supervision during this research and pay my gratitude to Prof. Dr. Thomas Dandekar for his generous support. I also thanks to my beloved wife and colleague Mrs. Saman Majeed (Doctoral Researcher) for her support during this research, development and technical documentation.

References

[1] Windchill, Reviewed 06 November2008<http://www.ptc.com/WCMS/files/56909/en/2757_Windchill_bro_ViewONLY.pdf>

[2] CIM Products, last reviewed, 01 October 2009, <http://www.contact.de/pdm-plm-products>

[3] Z. Ahmed, S. K. Ganti, H. Kyhlbäck: "Design Artifact's, Design Principles, Problems, Goals and Importance ", In Proceedings 4th International Statistical Conference May 9-11, Paper ID 42, Vol. 15, 57-68, ISBN 978-969-8858-04-9, 2008

[4] Klemmer, S. R and Lee, B: "Notebooks that Share and Walls that Remember: Electronic Capture of Design Education Artifacts". In Conference Supplement to UIST, 2005: ACM Symposium on User Interface Software and Technology. October 23-26, 2005, Seattle, WA

[5] Z. Ahmed: "Proposing LT based Search in PDM Systems for Better Information Retrieval", Category: Original Research Paper, International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004), Volume 1, Issue 4, P86-100, December 2010

[6] Z. Ahmed, "Contributions to advance Product Data Management Systems (PDMs): Towards Flexible Graphical User Interface and Semantic Oriented Search for Web based PDMs", Chapter 2, ISBN: 9783838324951, LAP Lambert Academic Publishing Germany, March 2010.

[7] Z. Ahmed, D. Gerhard: "Contributions of PDM Systems in Organizational Technical Data Management", Research Paper, Published in the proceedings of The First IEEE International Conference On Computer, Control & Communication (IEEE-IC4 2007), 12-13 November 2007

Supplementary Web Links

1. Flex, reviwed 20 August 2010, http://www.adobe.com/products/flex/?promoid=BPDEP
2. AJAX, reviwed 20 August 2010, http://articles.sitepoint.com/article/build-your-own-ajax-web-apps
3. OpenLaszlo, reviwed 20 August 2010, http://www.openlaszlo.org/
4. Silverlight, reviwed 20 August 2010, http://www.silverlight.net/

Author Biographies

Zeeshan Ahmed; (born 15.01.1983) a Software Research Engineer by profession and presently working in the Department of Bioinformatics Biocenter University of Wuerzburg Germany. He has on record more than 12 years of University Education and more than 8 years of professional experience of working within different multinational organizations in the field of Computer Science with emphasis on software engineering of product line architecture based artificially intelligent systems.

Process Orchestration for Intrusion Detection System based on SOA and Event Driven Architecture Principles

K.V.S.N.Rama Rao [1] , Pandu Prudhvi [2] , Manas Ranjan Patra[3]
[1]BSIT , Hyderabad, India
[2]TechMahindra, Hyderabad, India
[3]Berhampur University, Berhampur, Orissa
{kvsnramarao@yahoo.co.in , pswamy.2009@gmail.com, mrpatra12@gmail.com }

Abstract: *As the dependency on the internet in the recent years is increasing greatly, threats and vulnerabilities were also rising in sync to it. Several security systems like Intrusion detection systems were developed to battle against these threats. But the existing intrusion detection systems were not succeeding against these threats, as they are unable to address challenges that surround different types of attacks. These systems are designed to deliver the best performance but not able to deal with some attacks because they lack service oriented architecture to support increasingly diverse clients with various network and device capabilities. It is evident that no single technique can guarantee protection against future attacks. .Hence there is a need for integrated architecture which can provide robust protection against a complete spectrum of threats. In this paper, we propose a SOA based architecture model for IDS and its process orchestration based event driven architecture.*

Keywords: *IDS, vulnerability, SOA, EDA, web services, architecture.*

1. Introduction

With the growing use of Internet, attackers are becoming active in identifying the flaws in operating systems, underlying network protocols, and different software implementations. They are able to make sophisticated attacks on information resources. As a defense it is most common to use host based solutions like antivirus software, fire walls etc. These approaches have drawbacks in being insufficiently fast to meet new threats. Now a day due to globalization, multiple stake holders are involving in the activities of any organization. For example, in the case of IT projects several stake holders like end users, customers, vendors, legal entities and many others are involved to complete the project successfully.

In such a distributed and heterogeneous setup, security policies and their implementations suffer from the inability to cope with the flexibility of multi-site and multi-organization rules and the rigidity of a strong de-militarized zone. In this paper, we discuss the service oriented approach to build intrusion detection systems. This approach has the following key features:

i) Helps in identifying and analyzing the tasks performed by an IDS at a higher-level of abstraction.

ii) Helps in designing and building independently scalable components to deal with different aspects of an attack scenario.

iii) Helps in modeling the interactions among these components in an efficient and flexible manner

iv) Helps in adding new services when necessary.

2. Intrusion Detection

Intrusion [1] is defined as set of actions aimed to compromise the security goals. Intrusion Detection is the process of identifying and responding to intrusion activities. While modeling IDS we assume that normal and intrusive activities have distinct evidence and the system activities are observable. Any IDS have few important components [1].

a)Sensor or Agent: It Monitors and analyze network activity

b) Detection Engine: It contains rules and various detection models.

c) Decision Engine: On receiving alarm from detection engine it takes appropriate action and generates report.

Any typical IDS will focus on three areas of detection methodologies.

a) Signature based detection: A signature [1] is a pattern that corresponds to a known threat. Signature-based detection is the process of comparing signatures against observed events to identify possible incidents, Examples of signatures are: A telnet attempt with a username of “root”, which is a violation of an organization’s security policy, an e-mail with a subject of “Free pictures!”, and an attachment filename of “freepics.exe”, which are characteristics of a known form of malware. Signature-based detection cannot track and understand the state of complex communications, so it cannot detect most attacks that comprise multiple events.

Hackers [2] often attack networks through tried and tested methods from previously successful assaults. These attacks have been analyzed by network security vendors and a detailed profile, or attack signature, has been created. Signature detection techniques identify network assaults by looking for the attack fingerprint within network traffic and matching against an internal database of known threats. Once an attack signature is identified, the security system delivers an attack response, in most cases a simple alarm or alert.Success in preventing these attacks depends on an up-to-the-minute database of attack signatures,compiled from previous strikes. The drawback to systems that rely mainly, or only, on signature detection is clear: they can only detect attacks for which there is a released signature. If signature detection techniques are employed in isolation to protect networks, infrastructure remains vulnerable to any variants of known signatures, first-strike attacks, and Denial of Service attacks.

b) Anomaly-based detection: It compares definitions [1] of what activity is considered normal against observed events to identify significant deviations. This method uses profiles that are developed by monitoring the characteristics of typical activity over a period of time. The IDS then compares the characteristics of current activity to thresholds related to the profile. Anomaly-based detection methods can be very effective at detecting previously unknown threats. Common problems with anomaly-based detection are inadvertently including malicious activity within a profile, establishing profiles that are not sufficiently complex to reflect real-world computing activity, and generating many false positives.

Anomaly detection [2] techniques are required when hackers discover new security weaknesses and rush to exploit the new vulnerability. When this happens there are no existing attack signatures. The Code Red virus is an example of a new attack, or first strike, which could not be detected through an available signature. In order to identify these first strikes, IDS products can use anomaly detection techniques, where network traffic is compared against a baseline to identify abnormal—and potentially harmful—behavior. These anomaly techniques are looking for statistical abnormalities in the data traffic, as well as protocol ambiguities and atypical application activity. Today's IDS products do not generally provide enough specific anomaly information to prevent sophisticated attacks and if used in isolation, anomaly detection techniques can miss attacks that are only identifiable through signature detection.

c) Denial of Service (DoS) Detection [2]

The objective of DoS and Distributed DoS attacks is to deny legitimate users access to critical network services. Hackers achieve this by launching attacks that consume excessive network bandwidth or host processing cycles or other network infrastructure resources. DoS attacks have caused some of the world's biggest brands to disappoint customers and investors as Web sites became inaccessible to customers, partners, and users—sometimes for up to twenty-four hours. IDS products often compare current traffic behavior with acceptable normal behavior to detect DoS attacks, where normal traffic is characterized by a set of pre-programmed thresholds. This can lead to false alarms or attacks being missed because the attack traffic is below the configured threshold.

IDS can be deployed at the following places to monitor activities.

a) Host based IDS: which monitors the characteristics of a single host and the events occurring within that host for suspicious activity.
Ex: Analyze shell commands, Analyze system calls made by send mails etc.

b) Network-Based IDS: this monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity.
Ex: Watch for violations of protocols and unusual connection patterns, look into the data portions of the packets for malicious command sequences etc.

In order to robustly protect enterprise network against the complete spectrum of threats and vulnerabilities, there is a need for robust architecture. But due to the lack of superior architectural support, current IDS are facing various challenges which are discussed below.

3. Current IDS Challenges

Intrusion Detection Systems today are facing several challenges [2].

Incomplete attack coverage: IDS products typically focus on Signature, Anomaly, or Denial of Service detection. Network security managers have to purchase and integrate point solutions from separate vendors or leave networks vulnerable to attack.

Inaccurate detection: IDS products' detection capabilities can be characterized in terms of accuracy and specificity. Accuracy is often measured in true detection rate—sometimes referred to as the false negative rate—and the false-positive rate. The true detection rate specifies how successful a system is in detecting attacks when they happen. The false-positive rate tells us the likelihood that a system will misidentify benign activity as attacks. Specificity is a measure of how much detailed information about an attack is discovered when it is detected. IDS products today are lacking in both accuracy and specificity and generate too many false-positives, alerting security engineers of attacks, when nothing malicious is taking place. In some cases, IDS products have delivered tens of thousands of false-positive alerts a day. There is nothing more corrosive to network vigilance than a jumpy security system, which is continually issuing false alarms.

Detection, not prevention: Systems concentrate on attack detection. Preventing attacks is a reactive activity, often too late to thwart the intrusion.

Designed primarily for sub-100Mb/s networks: Solutions have simply not kept up with the speed and sophistication of network infrastructure and cannot accurately monitor higher-speed or switched networks.

Performance challenged: Software applications running on general purpose PC/server hardware do not have the processing power required to perform thorough analysis. These underpowered products result in inaccurate detection and packet dropping, even on low bandwidth networks.

Lack of high-availability deployment: Single port products are not able to monitor asymmetric traffic flows. Also, with networks becoming a primary mechanism to interact with customers and partners, forward-thinking organizations have developed back-up systems should their current infrastructure fail in any way. The inability of current IDS products to cope with server failovers renders them virtually useless for any mission-critical network deployment.

Poor scalability: Primarily designed for low-end deployments, today's IDS products do not scale for medium and large enterprise or government networks. Here monitored bandwidth, the number of network segments monitored, the number of sensors needed, alarm rates, and the geographical spread of the network exceed system limits.

No multiple policy enforcement: Current products generally support the selection of only one security policy for the entire system, even though the product may monitor traffic belonging to multiple administrative domains—in an enterprise this could be the finance, marketing, or HR functions. This one size fits all approach is no longer acceptable for organizations that require different security policies for each function, business unit, or geography.

Require significant IT resources: IDS products today require substantial hands-on management—for example,the simple task of frequent signature updates can take up a lot of time and skilled engineering resources,delivering a very high total cost of ownership.
Concisely, one can state that many of the IDS implementations are not designed to co-operate. To address these challenges, a new architecture needs to be developed for even the most demanding enterprise networks.

Hence we propose an architecture that works for problem of a multi-site IDS for a multi-business scenario, shown in Figure.1, where each business can have a custom defined set of rules implemented at each location of choice.

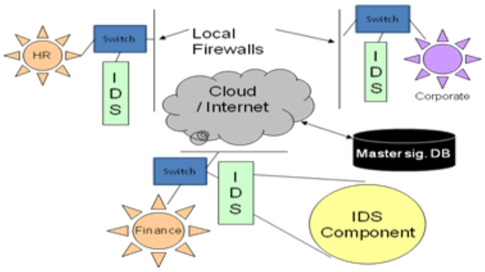


Figure 1: Distributed IDS deployed across mutli-location corporate network

4. Model Architecture for Proposed IDS

Figure 2 summarizes the architecture of the fast Ethernet IDS system designed.

The proposed solution contains a Cache component that collects network packets.

The functionality of the components are explained below.

Sampler: The Sampler randomly/heuristically picks up sample packet windows (series of contiguous packets) and sends them to the Network Packet Analyzer component. The sampling can be done in a random fashion or by using a heuristic.

Network Packet Analyzer: The Analyzer and the Pre-processing engine analyze the packets and convert them into a standard XML format by stripping the network and DLL headers. This metadata is sent for processing to the next component i.e. the "Rules Engine" which can be an SOA component.

Business Rules Engine: The Rules engine is a SOA [3] enabled component of the application that facilitates the XML packet to be checked for anomalies against suspicious activities and pre-defined business rules. This component should be able to detect packets from invalid/untrusted IPs and domains. DoS attacks, Filtering, Screening, Authentication, Trust, etc. related issues can be addressed at this component. The Rules engine should be SOA enabled to allow the organization to implement and customize the rules based on the location of the IDS on the network. For example in a large enterprise, HR may need a different set of rules implemented as against the finance and there may be some organizational rules applicable to all departments. Rules must be classified as preemptive/non-preemptive. A web-service [4] client can allow for posting of rules to be consumed and for rules to be published [5] from one instance of the IDS to another which is one of the many advantages of a SOA enabled system. The rules engine upon detecting anomaly will automatically forward to alert agent component or manual intervention component.

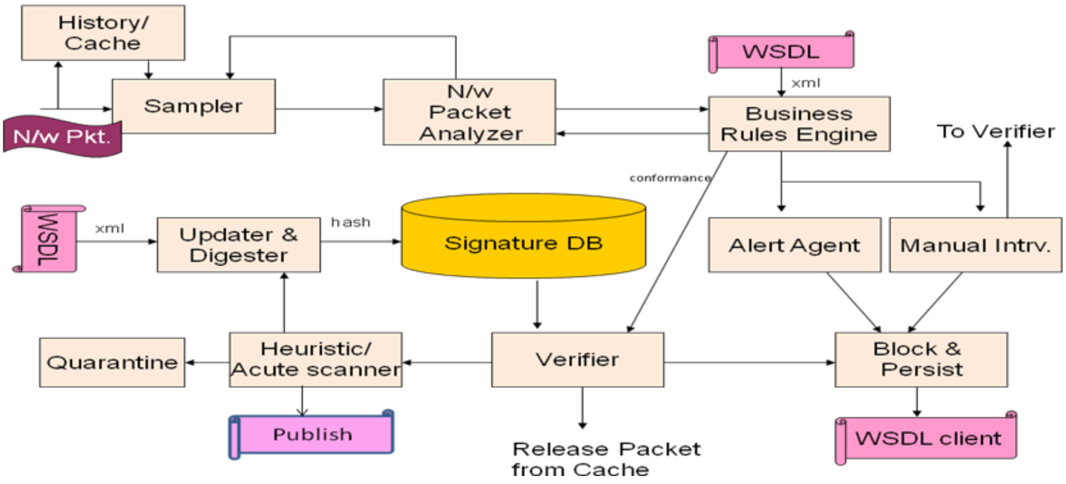


Figure 2: Proposed Model of the IDS component showing the SOA enabled external interfaces and custom built internal components

Alert Agent: If directed to alert agent component then the alerts are audited, logged, and mailed to concerned authorities.

Manual Intervention: On the other hand, if the threat is not that much harmful , they are directed to manual intervention component where they can be manually addressed by administrator of the location. The Manual intervention application may flag to either further analyze or release the associated packets to the verifier. If all packets in a sample packet window are cleared by Business Rules Engine, then the packets go for a check of known attack signatures to the verifier.

Verifier: The Verifier component checks the packets against attacks picked from a local signature database. This DB is pre-populated from external and publicly known signatures and other IDS instance detected signatures. These signatures are also batch mode synchronized between IDS instances through the Updater and Digester service component. Whenever the verifier component detects a known signature match, it immediately discards the packet and the payload. In case of innocuous packets it can inform the cache to release them. For packets that have matched a possible known attack, the packets and the payload can be sent into the heuristic and acute scanner. Since the signatures are hashed, comparing them in the verifier against new XMLs and network packet payloads becomes easy and quick to achieve the “fast Ethernet” speeds that this architecture claims.

Heuristic and Acute Scanner: This will perform further analysis to detect newer form of attacks or decisively declare a packet/source as safe. This can over a period of time detect new attacks and recover from false alarms.

Thus if a payload was marked as possibly harmful, a fuzzy logic AI agent running in the heuristic scanner can verify the safety or the hostility of the payload to a pre-determined degree of threshold (say 25% to 80%) before declaring and publishing it to other instances through the Master DB and also updating the local DB.

Updater and Digester: The updater listens for updates on a daily basis from the Master DB, which is connected on the cloud and sends web-service based publish notices to all instances. The updater then picks up these XMLs and their packet payloads and digests them using fast and compressive hashing algorithms that compact this information and store it in the local signature DB. The updater and digester component in conjunction with the Business Rules engine thus ensures that over a period of time the IDS learns to detect unknown attacks and thus can prevent them as well making it a true IDS

Master Database: It is kept updated through SOA components about attacks detected or false alarms nullified at the distributed locations. The Master Database on the next day updates all IDS instances local databases

Block and Persist: This component fires whenever the Manual intervention module marks a XML cum payload pair as suspicious or malicious or if the alerter escalates a known business rule violation. The component simply publishes the packet to be updated into the local DB via the Updater and Digester service and to the MasterDB which runs a similar Updater and Digester service

In the proposed model, several components such as Business Rules Engine, Block and persist, Heuristic and acute scanner and Updater & Digester are web services.

The advantage of using web-services clients here is that it becomes easy to update the remote DB and the local DB through common interfaces and in future to publish the same to other external service consumers as well, e.g. security provider Databases or public signature databases – on which the current system relies as well.

The best architecture to integrate web services is known to be service oriented architecture (SOA).A brief description of web service, SOA, SOAD Process are discussed in the next section.

5. Service Oriented Architecture

Service Oriented Architecture (SOA) [6] is a business-centric IT architectural approach that supports integrating your business as linked, repeatable business tasks, or services.

A service is a mechanism to enable access to one or more capabilities, where the access is provided using a prescribed interface and is exercised consistent with constraints and policies as specified by the service description. A Web service provides one way of implementing the automated aspects of a given business or technical service.

Services generally adhere to the principles of service-orientation such as abstraction, autonomy, composability, discoverability, formal contract, loose coupling, reusability, statelessness

Why SOA?

SOA helps create greater alignment between IT and line of business while generating more flexibility - IT flexibility to support greater business flexibility. Your business processes are changing faster and faster and global competition requires the flexibility that SOA can provide. SOA can help you get better reuse out of your existing IT investments as well as the new services you're developing today. SOA makes integration of your IT investments easier by making use of well-defined interfaces between services. SOA also provides an architectural model for integrating business partners', customers' and suppliers' services into an enterprise's business processes. This reduces cost and improves customer satisfaction.

SOA is a suitable architecture style when reusability, integration and agility are key concerns for an enterprise.

Basically the four tenets of Service orientation [7] are as follows

- Boundaries are explicit
- Services are autonomous
- Services share schema and contract, not class
- Compatibility is based upon policy

SOA Design Principles

- Deciding what functionality makes sense to expose as a service
- Separating and modularizing the business logic to facilitate reuse and flexibility
- Loosely coupling services to support rapid development when requirements change
- Designing an appropriate granularity of services
- Planning and implementing all the SOAD steps.

5.1 SOAD Process [8]:

The term Process means "sequence of steps required to develop or maintain software"[9].A process deals with what of developing a software while a methodology deals with how of developing a software. With the introduction of object oriented paradigm, OOAD process has been in use extensively.

Object oriented analysis and design process involves modeling real world objects based on the requirements described as a set of use cases, realizing the use cases through a process of identifying the analysis classes(boundary, control and entity) and mapping the analysis to technology elements that constitute the design classes. Classes are fine grained elements that are tightly coupled. Design classes can be implemented through programming and tested to develop the required application. But OOAD Process presents several difficulties [10].Since the OO applications granularity is at class level, there will be tight coupling and strong associations because class hierarchies are based on inheritance. .But on the other hand, services are loosely coupled.

In Service model, there will two important roles. Service Provider who exposes services and Service Consumer who consumes service. There will be a service contract between these two parties to define the type of messages they can exchange or operations they can perform. Also there will be a data contract between the client and the service. These contracts enables loose coupling. The key considerations of service model such as reusability, integration and agility will result in four types of services [11].

Client services: These deliver content to the business users that require an aggregated enterprise view. They provide presentation content to the “front-end” applications of the enterprise such as Portal, dashboard or CRM applications that provide the necessary presentation capabilities and typically are service consumers for the other services in the Enterprise.

Business Process management services (Process services): These allow for externalization of business processes in an orchestrable fashion resulting in agility for the enterprise.

Business Application services (Activity services): These are reusable business level services that can be orchestrated as part of a configured business process.

Data Services (Entity services): These encapsulate access to data in various sources such as ERP, legacy, a data warehouse or a system external to the organizational context

These four services are integrated by Enterprise service bus. Considering these 4 services, SOAD process consists the following steps.

Step1: Gather objectives and business requirements of the application.

Step2: Perform Business Process Modeling (BPM) that involves identification of business processes and workflows that applications in the enterprise would need to support to meet the business objectives. The business process model provides the workflows that may be expressed as Business Process Execution Language(BPEL), configured and orchestrated to generate the Business process services. The business process model generated through BPEL is the key artifact of SOAD Process. This will serve as an input to develop four services which were discussed above.

Step3: Implementation: A technology stack is chosen for implementation of services.

The scope of this paper covers the first two steps of SOAD process. In the following sections, we present Business process modeling generated through Business Process Execution Language (BPEL) for the IDS model architecture described in section 4.BPEL generates work flows and process orchestrations.

6. Process Modeling For IDS Using BPEL:

The process modeling will be done for the architecture that is explained in section 4. In that architecture several components are web services.

Hence the next task in SOAD process is to perform Business Process Modeling using BPEL which expresses the workflows.

Process Orchestration diagram is shown in figure 3.

The corresponding BPEL code given below.

<u>IDSProcess.bpel</u>

```
<process name="IDSProcess"

targetNamespace="http://ids.security.com/IDS/idsServices/IDSProcess"

xmlns="http://schemas.xmlsoap.org/ws/2003/03/business-process/"

xmlns:client="http://ids.security.com/IDS/idsServices/IDSProcess"

xmlns:ora="http://schemas.oracle.com/xpath/extension"

xmlns:bpelx="http://schemas.oracle.com/bpel/extension"

xmlns:bpws="http://schemas.xmlsoap.org/ws/2003/03/business-process/"

xmlns:ns1="http://oracle.com/sca/soapservice/Application1/Project1/NetWrokPacketAnalyser"

xmlns:ns2="http://xmlns.oracle.com/pcbpel/adapter/jms/Application1/Project1/AlertManagerService"

xmlns:task="http://xmlns.oracle.com/bpel/workflow/task"

xmlns:taskservice="http://xmlns.oracle.com/bpel/workflow/taskService"

xmlns:wfcommon="http://xmlns.oracle.com/bpel/workflow/common"

xmlns:ns3="http://oracle.com/sca/soapservice/Application1/Project1/VerifierService"

xmlns:ns4="http://oracle.com/sca/soapservice/Application1/Project1/DigesterService"
```

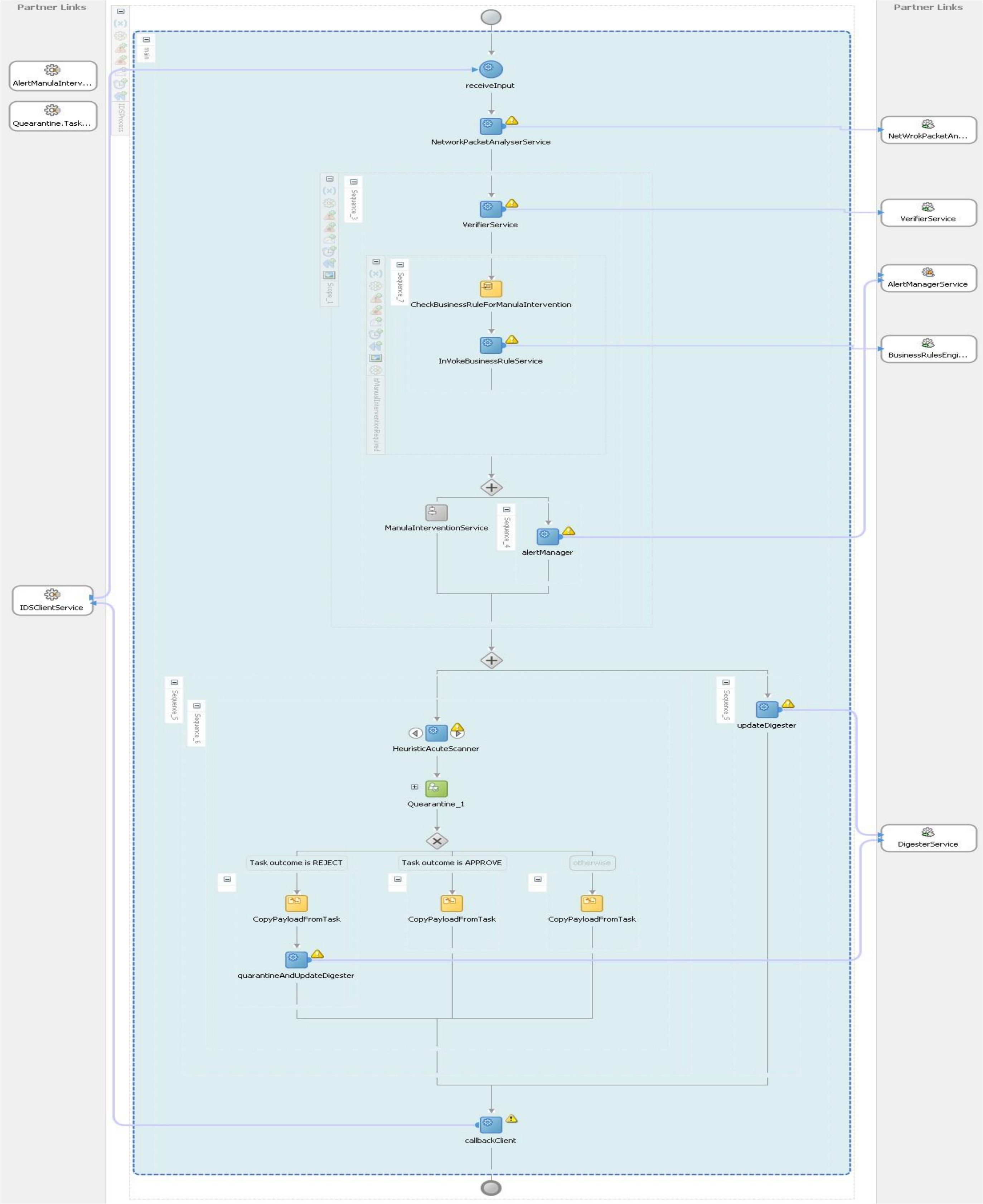


Figure 3: Business Process modeling for IDS

```
xmlns:ns5="http://oracle.com/sca/soapservice/Application1
/Project1/BusinessRulesEngineService">

   <!—<!--
     PARTNERLINKS
       <!--
     The 'client' role represents the requester of this service.
It is   used for callback. The location and correlation
information associated      with the client role are
automatically set using WS-Addressing. - - >

Link myRole="execute_ptt"
name="NetWorkPacketAnalyser"

partnerLinkType="ns1:NetWorkPacketAnalyser"/>
       <partnerLink myRole="alertManager_role"
name="AlertManagerService"
                  partnerLinkType="ns2:alertManager_plt"/>
  -->
    <partnerLinks>
```

Request from IDS client service are received as input by a proxy server. The proxy manager will select a sample of packets and directs them to Network packet analyzer through sendPayload() method. The Network Analyzer will connect to its partner links. and convert the packets into a standard XML format by stripping the network and DLL headers. This Meta data is sent for validation against business rules.

The Rules engine is SOA enabled to allow the organization to implement and customize the rules based on the location of the IDS on the network. For example in a large enterprise, HR may need a different set of rules implemented as against the finance and there may be some organizational rules applicable to all departments. A web-service client can allow for posting of rules to be consumed and for rules to be published from one instance of the IDS to another which is one of the many advantages of a SOA enabled system.

The rules engine upon detecting anomaly will automatically forward to alert agent component or manual intervention component.

The orchestration logic is represented below

```
<sequence name="main">
       <!-- Receive input from requestor. (Note: This
maps to operation defined in IDSProcess.wsdl) -->
       <receive                    name="receiveInput"
partnerLink="IDSClientService"
portType="client:IDSProcess"    operation="process"
variable="inputVariable" createInstance="yes"/>

       <!--
         Asynchronous callback to the requester. (Note:
the callback location and correlation id is
transparently handled using WS-addressing.)
       -->
  <invoke     name="NetworkPacketAnalyserService"
partnerLink="NetWrokPacketAnalyser"/>
       <scope name="Scope_1">
          <bpelx:annotation>
             <bpelx:general>
                <bpelx:property
name="userLabel">Business                     Rules
</bpelx:property>
             </bpelx:general>
          </bpelx:annotation>
          <sequence>
             <invoke              name="VerifierService"
partnerLink="VerifierService"/>
             <scope
name="isManualInterventionRequired">
                <bpelx:annotation>
                   <bpelx:pattern
patternName="bpelx:decide"></bpelx:pattern>
                </bpelx:annotation>
                <sequence>
                   <bpelx:checkpoint
name="CheckBusinessRuleForManulaIntervention"/
>
                   <invoke
name="InVokeBusinessRuleService"

partnerLink="BusinessRulesEngineService"/>
                </sequence>
             </scope>
             <flow name="Rules">
                <sequence>
                   <invoke name="alertManager"

partnerLink="AlertManagerService"/>
                </sequence>
```

The high level design of Rules engine service is shown in the Figure 4.

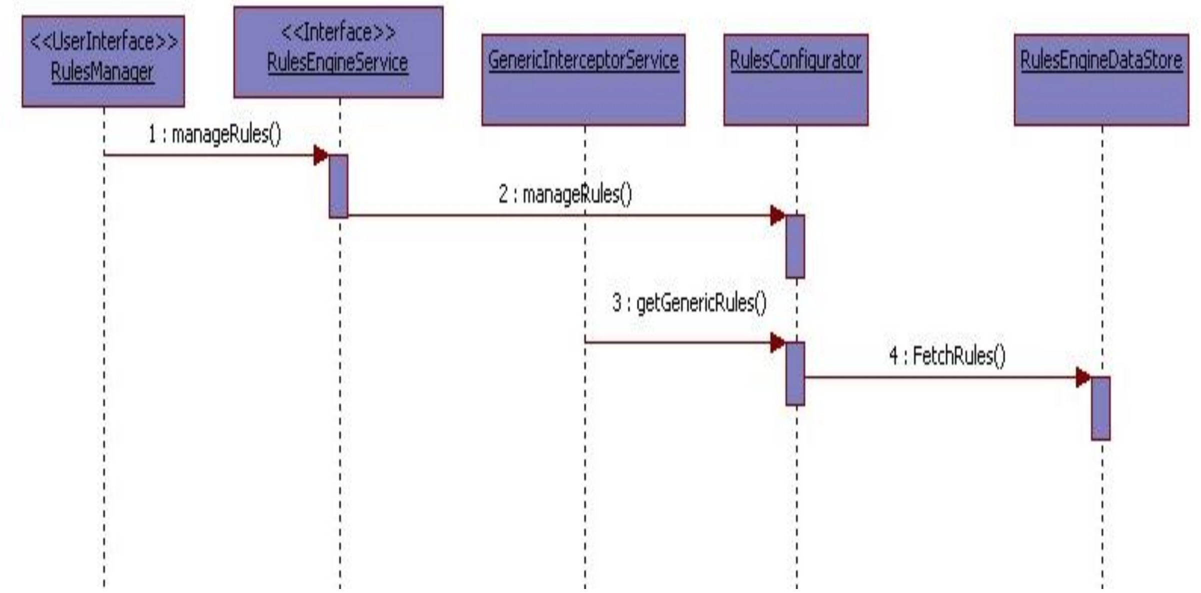


Figure 4: Rules Engine Service

The Rules Manager Interface will contact the Rules Engine Service Interface by using the method manageRules(). This service will invoke Rules configarator. Concurrently another service called Generic Interceptor Service will request for generic rules from the RulesConfigarator through getGenericRules() method. RulesConfigarator will in turn contact RulesEngineDataStore through fetchRules() method and will fetch both generic and specific rules. Hence by using this service, each administrative domain in an enterprise like HR, Marketing etc can customize their own business rules in conjunction with enterprise wide business rules. So the use of this service has conquered one of the challenges of IDS.

After passing the business rules check, the pay load will be directed to verifier for second level of checking. If this check is successful and found that the packet is not harmful, packet will be released otherwise it will be sent for heuristic and acute scanner.

The corresponding orchestration logic is presented below.

```
<sequence name="ManulaInterventionService">
        <sequence>
                <scope
name="AlertManulaInterventionToVerifier_1"

xmlns="http://schemas.xmlsoap.org/ws/2003/03/busi
ness-process/"

xmlns:wf="http://schemas.oracle.com/bpel/extension/
workflow"

wf:key="AlertManulaInterventionToVerifier_1_glob
alVariable">
                    <bpelx:annotation
xmlns:bpelx="http://schemas.oracle.com/bpel/extensi
on">
                      <bpelx:pattern
patternName="bpelx:workflow"></bpelx:pattern>
                    </bpelx:annotation>
                    <variables>
                      <variable
name="initiateTaskInput"

messageType="taskservice:initiateTaskMessage"/>
```

```
                  <variable
name="initiateTaskResponseMessage"

messageType="taskservice:initiateTaskResponseMes
sage"/>
                </variables>
                <sequence>
                  <assign
name="AlertManulaInterventionToVerifier_1_Assig
nTaskAttributes">
                    <copy>
                      <from
expression="number(3)"/>
                      <to
variable="initiateTaskInput"
                        part="payload"

query="/taskservice:initiateTask/task:task/task:priorit
y"/>
                    </copy>
                    <copy>
                      <from>
                        <payload
xmlns="http://xmlns.oracle.com/bpel/workflow/task"
/>
                      </from>
                      <to
variable="initiateTaskInput"
                        part="payload"

query="/taskservice:initiateTask/task:task/task:payloa
d"/>
                    </copy>
                  </assign>
                  <invoke
name="initiateTask_AlertManulaInterventionToVerif
ier_1"

partnerLink="AlertManulaInterventionToVerifier.Ta
skService_1"

portType="taskservice:TaskService"
                      operation="initiateTask"

inputVariable="initiateTaskInput"

outputVariable="initiateTaskResponseMessage"/>
                  <receive
name="receiveCompletedTask_AlertManulaIntervent
ionToVerifier_1"

partnerLink="AlertManulaInterventionToVerifier.Ta
skService_1"

portType="taskservice:TaskServiceCallback"
operation="onTaskCompleted"

variable="AlertManulaInterventionToVerifier_1_glo
balVariable"
                      createInstance="no"/>
                </sequence>
              </scope>
              <switch
name="humanIntervention">
                <case
condition="bpws:getVariableData('AlertManulaInter
ventionToVerifier_1_globalVariable', 'payload',
'/task:task/task:systemAttributes/task:state') =
'COMPLETED' and
bpws:getVariableData('AlertManulaInterventionToV
erifier_1_globalVariable', 'payload',
'/task:task/task:systemAttributes/task:outcome') =
'APPROVE'">
                  <bpelx:annotation>
                    <bpelx:pattern>Task
outcome is APPROVE</bpelx:pattern>
                    <bpelx:general>
                      <bpelx:property
name="userLabel">Task

outcome
                                      is

APPROVE</bpelx:property>
                    </bpelx:general>
                  </bpelx:annotation>
                  <sequence>
                    <assign/>
                    <invoke
name="alertManager"

partnerLink="AlertManagerService"/>
                  </sequence>
```

The high level design of verifier service is shown in figure 5.

In this service, the incoming pay load signature is checked with the pre-populated database from external and publicly known signatures and other IDS instance detected signatures. This will be achieved by using signatureVerification() method .

The updater Service listens for updates on a daily basis from the Master DB, which is connected on the cloud and sends web-service based publish notices to all instances. The updater then picks up these XMLs and their packet payloads and digests them using fast and compressive hashing algorithms that compact this information and store it in the local signature DB as signatures.

So by using this service, one more challenge of IDS can be conquered such as frequent signature updates taking up a lot of time and skilled engineering resources, delivering a very high total cost of ownership.

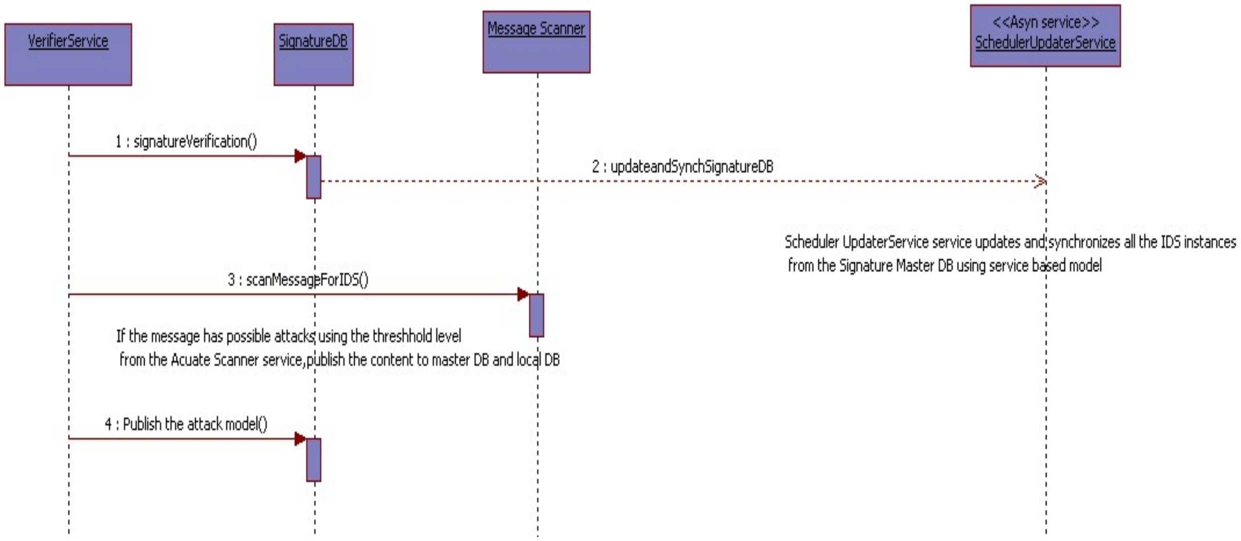

VerifierService
SignatureDB
Message Scanner
<<Asyn service>>
SchedulerUpdaterService
1 : signatureVerification()
2 : updateandSynchSignatureDB
Scheduler UpdaterService service updates and synchronizes all the IDS instances
from the Signature Master DB using service based model
3 : scanMessageForIDS()
If the message has possible attacks using the threshhold level
from the Acuate Scanner service,publish the content to master DB and local DB
4 : Publish the attack model()

Figure 5: Verifier Service

Another functionality of this service is scanning messages for possible attacks. As already mentioned the verifier checks the hash of the XML and the payload from the DB against the incoming XML and payload of the sample window, which enables to detect a match for a possible harmful packet. For packets that have matched a possible known attack, the packets and the payload can be sent into the heuristic and acute scanner that can perform further analysis to detect newer form of attacks or decisively declare a packet/source as safe. This can over a period of time detect new attacks and recover from false alarms. Thus one more challenge of IDS such as lacking in both accuracy and specificity and generate too many false alarms is conquered.

Thus if a payload was marked as possibly harmful, a fuzzy logic AI agent running in the heuristic scanner can verify the safety or the hostility of the payload to a pre-determined degree of threshold (say 25% to 80%) before declaring and publishing it to other instances through the Master DB and also updating the local DB.

The Orchestration logic for Heuristics scanner and digester is presented below

```
  <sequence name="Sequence_5">
          <invoke name="updateDigester"
partnerLink="DigesterService"/>
        </sequence>
        <sequence name="Sequence_5">
          <sequence>
            <invoke
name="HeuristicAcuteScanner"/>
            <scope name="Quearantine_1"

xmlns="http://schemas.xmlsoap.org/ws/2003/03/busi
ness-process/"

xmlns:wf="http://schemas.oracle.com/bpel/extension/
workflow"

wf:key="Quearantine_1_globalVariable">
              <bpelx:annotation
xmlns:bpelx="http://schemas.oracle.com/bpel/extensi
on">
                <bpelx:pattern
patternName="bpelx:workflow"></bpelx:pattern>
              </bpelx:annotation>
              <variables>
                <variable name="initiateTaskInput"

messageType="taskservice:initiateTaskMessage"/>
                <variable
name="initiateTaskResponseMessage"
```

```
messageType="taskservice:initiateTaskResponseMes
sage"/>
            </variables>
            <sequence>
              <assign
name="Quearantine_1_AssignTaskAttributes">
                <copy>
                  <from
expression="number(3)"/>
                  <to
variable="initiateTaskInput"
                    part="payload"

query="/taskservice:initiateTask/task:task/task:priorit
y"/>
                </copy>
                <copy>
                  <from>
                    <payload
xmlns="http://xmlns.oracle.com/bpel/workflow/task"
/>
                  </from>
                  <to
variable="initiateTaskInput"
                    part="payload"

query="/taskservice:initiateTask/task:task/task:payloa
d"/>
                </copy>
              </assign>
              <invoke
name="initiateTask_Quearantine_1"

partnerLink="Quearantine.TaskService_1"

portType="taskservice:TaskService"
                  operation="initiateTask"

inputVariable="initiateTaskInput"

outputVariable="initiateTaskResponseMessage"/>
              <receive
name="receiveCompletedTask_Quearantine_1"

partnerLink="Quearantine.TaskService_1"

portType="taskservice:TaskServiceCallback"

operation="onTaskCompleted"

variable="Quearantine_1_globalVariable"
                  createInstance="no"/>
            </sequence>
          </scope>
          <switch name="taskSwitch">
            <case
condition="bpws:getVariableData('Quearantine_1_gl
obalVariable', 'payload',
'/task:task/task:systemAttributes/task:state') =
'COMPLETED' and
bpws:getVariableData('Quearantine_1_globalVariabl
e', 'payload',
'/task:task/task:systemAttributes/task:outcome') =
'REJECT'">
              <bpelx:annotation>
                <bpelx:pattern>Task outcome is
REJECT</bpelx:pattern>
                <bpelx:general>
                  <bpelx:property
name="userLabel">Task
                                  outcome
is

REJECT</bpelx:property>
                </bpelx:general>
              </bpelx:annotation>
              <sequence>
                <assign/>
                <invoke

partnerLink="DigesterService"

name="quarantineAndUpdateDigester"/>
              </sequence>
            </case>
            <case
condition="bpws:getVariableData('Quearantine_1_gl
obalVariable', 'payload',
'/task:task/task:systemAttributes/task:state') =
'COMPLETED' and
bpws:getVariableData('Quearantine_1_globalVariabl
e', 'payload',
'/task:task/task:systemAttributes/task:outcome') =
'APPROVE'">
              <bpelx:annotation>
                <bpelx:pattern>Task outcome is
APPROVE</bpelx:pattern>
                <bpelx:general>
                  <bpelx:property
name="userLabel">Task
outcome is
APPROVE</bpelx:property>
                </bpelx:general>
              </bpelx:annotation>
              <sequence>
                <assign/>
              </sequence>
```

All the period, the Master Database is kept updated through SOA components about attacks detected or false alarms nullified at the distributed locations. The Master Database on the next day updates all IDS instances local databases. So this makes the IDS as a true IDPS (Intrusion Detection and prevention system) because if the attack is detected at one location, the attack model is published to all other locations in the enterprise through a service.

Hence this proposed model has conquered one more challenge of IDS such as Current IDS are concentrating on detection, but not on prevention

6.1 Sub Processes Identification

As specified in [8], the business process model generated is the key artifact in SOAD process and serves as input to four sub processes. There are four sub processes.

1) Activity Services
2) Business Process services
3) Client Services
4) Data Services

To develop any of the 4 services discussed above, a series of generic steps.

- Service Identification.
- Analysis and design.
- Technology selection.
- Coding and Testing.
- Integrating services.

In the context of our IDS model, we have identified the 4 services as follows.

Activity Services: These are the applications that support our IDS tool activities. These applications not only support current state objectives but also meet the future state objectives. For example Different Algorithms (Pattern matching, Genetic, Intelligent), log generators, graph generators are identified as Activity services.

Business Process Services: Business Process model workflows are expressed as BPEL, configured and orchestrated to generate business process services. In our IDS model Business Rules, Heuristic & acute scanner, Updater & Digester are some of the Business Process services.

Client Services: These services facilitate the delivery of content through various channels such as web, mobile etc. For our IDS model a User Interface to customize business rules, displaying existing rules, capturing values in the form fields, conversion policy of converting a string into numerical value may be some of the examples of client services.

Data Services: These services define the way to store and access core data of the enterprise. For our IDS model adding new signatures to database, converting data to normalized structure and ensuring that database is maintained in clustered structure, access to master database are few examples of data services.

The SOAD process diagram along with four services are shown in figure 6.

For any IDS, several algorithms such as pattern matching algorithms, Dos detection algorithms etc are necessary for defense against severe threats. Activity services will maintain such algorithms and retrieve appropriate algorithm during the flow of events. Also they maintain log generators which will record all the activities of an attack. Graph generators will use this data and generate attack graphs. On the other hand, the client services will deal with presentation of an user interface for entry, display of rules and any conversion policies. The conversion policies are necessary as the host and network packet orders are different. The data services such as data normalization, cluster maintenance and interactions between local and master databases will aid us in maintaining the database in normalized form.

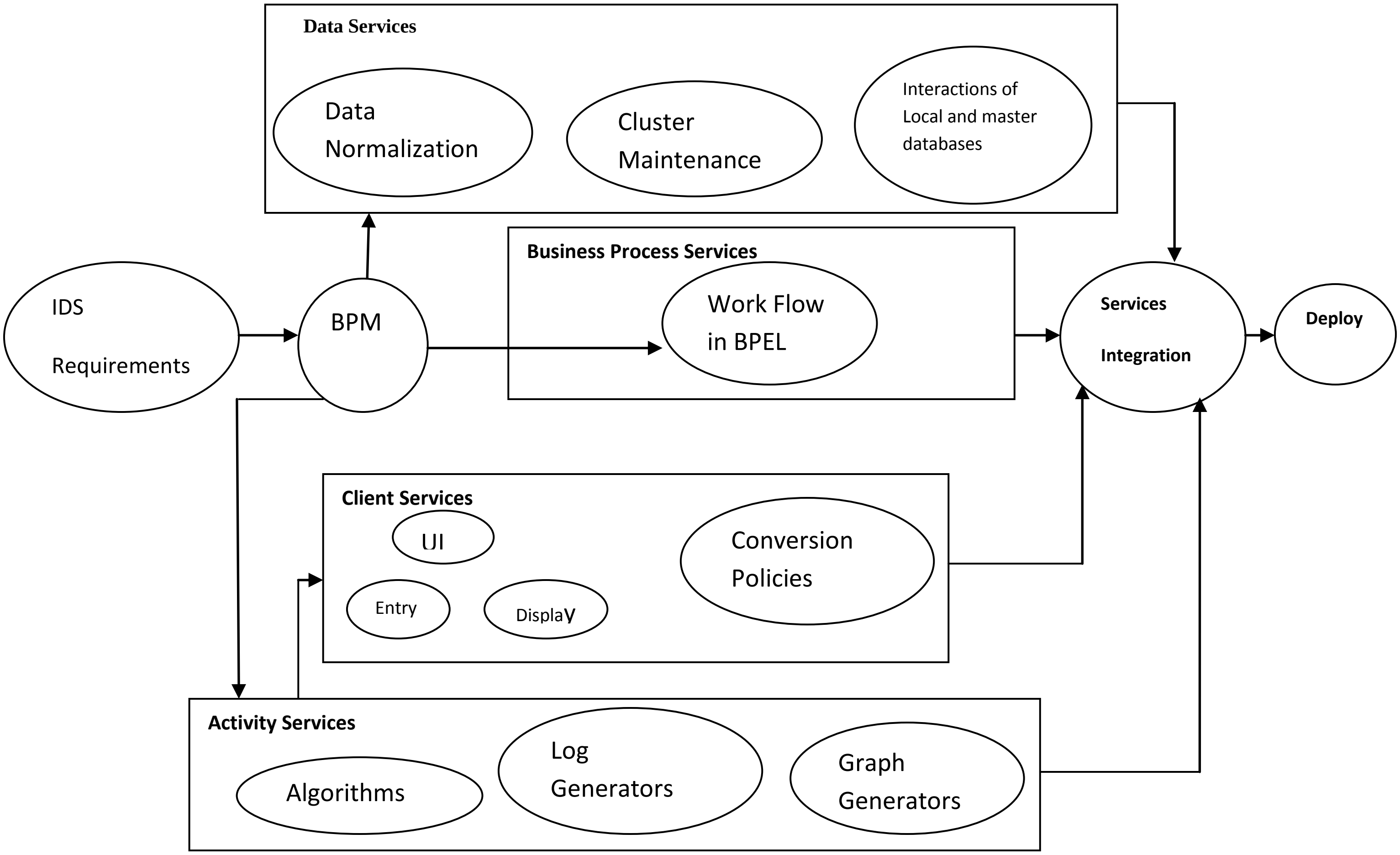


Figure 6: Sub-Processes Interaction

This will ensure that a rule will be present in the database only once. Also for efficient rule processing, the data is maintained in structured form such that related keywords are available in a single cluster. The important task of data service is to maintain the interactions with local database and master database. Synchronizing the data from several databases will be critical. The local database should update itself a new signature and inform the same to master database. The master database after a specified period of time, should update the local databases of other domains with all the new signatures. All these interactions between client, activity and data services will be synchronized by Business process services. The architecture is highly scalable and greatly interoperable.

7. RELATED WORK

In [12], based on danger theory, authors have proposed a four layer model of Immune based intrusion detection system. The first layer is danger sense layer which handles alert correlation problems, and to construct an intrusion scenario that would be detected by reacting to the balance of various types of alerts. The second layer is danger computation layer which calculates computes danger according to the intrusion alert 5-Tuple.The third layer is immune response layer which will detect the abnormal behavior. Fourth layer is spot disposal layer which will remove the dangerous behaviors. In [13], authors have proposed a Intrusion detection Intelligent agent system where several intelligent agents are integrated for providing in depth defense strategy against intrusions.

The main goals of this approach are its distributed architecture, scalability, efficiency and the use of intelligent agents. In [14], In order to enhance the availability and practicality of intelligent intrusion detection system based on machine learning in high-speed network, an improved fast inductive learning method for intrusion detection (FILMID) is designed and implemented. Accordingly, an efficient intrusion detection model based on FILMID algorithm is presented. In [15], a design scheme of intrusion detection system based on pattern matching algorithm is proposed. Also authors aimed at several key modules of intrusion detection system, a detailed analysis of data acquisition module, protocol processing module, feature matching module, log record module and intrusion response module is also given in this paper. Data acquisition module is responsible for capturing various types of hardware frames from network flow and handing these hardware frames to data pretreatment module and then the data pretreatment module strips off hardware frame heads and checks the integrity of messages. Based on application protocols hardware frames are sent to response protocol analyzing and processing modules respectively. For example, TELNET protocol has a process of packet. The pattern matching algorithm will judge for intrusion. If intrusion is found, alarm is given by intrusion response module and attack log is recorded. If no intrusion, it will make a detailed record of protocol operation log.

Conclusion

The proposed architecture and its design can manage the distributed system components efficiently. It allows new computing resources and services to be added dynamically. Most of the challenges faced by current IDS are addressed by the proposed architecture. Our future work aims at developing algorithms that would allow global distribution of various processing components.

References

[1] Rebecca Bace and Peter Mell "Intrusion Detection systems" NIST Special Publication on Intrusion Detection Systems
[2] McAfee network protection solutions "Next generation intrusion detection systems
[3]IBM Red book "Patterns: SOA Foundation Service Creation Scenario"
[4] Ali Arsanjani "How to identify, specify, and realize services for your SOA "
[5] Jim Amsden "Service realization"
[6] http://www-01.ibm.com/software/solutions/soa/
[7] Evdemon, J, 2005, "The Four Tenets of Service Orientation""http://www.bpminstitute.org/articles/article/article/the-four-tenets-of-service-orientation.html
[8]Shankar k," Service oriented analysis and design process for the enterprise"**,** 7th WSEAS International Conference on applied computer science, Venice, Italy, November 21-23, 2007,Pgs 366-371 , ISBN ~ ISSN:1750-5117 , 978-960-6766-18-3,ACM
[9] Humphrey, Watts. "A Discipline for Software Engineering. Reading", MA: Addison-Wesley Publishing Company, Inc., 1995.
[10]Zimmermann, O.; Krogdahl,P.; Gee, C., 2004, "Elements of Service-Oriented Analysis and Design".http://www.ibm.com/developerworks/webservices/library/ws-soad1/
[11]Kambhampaty, S,chandra, S. "Service Oriented Architecture for Enterprise Applications", WSEAS Transactions on Business and Economics. Issue 3, Volume 2, July 2005.
[12]Haidong Fu, Xiguo Yuan, Liping Hu," Design of a Four-layer Model Based on Danger Theory and AIS for IDS"2007,IEEE.
[13] Amine Berqia and Gustavo Nacsimento," A Distributed Approach For Intrusion Detection Systems"2004, IEEE**.**
[14] Wu Yang, Wei Wan , Lin Guo, Le-Jun Zhang," An Efficient Intrusion Detection Model Based On Fast Inductive Learning "Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007,IEEE.
[15] Zhang Hu," Design of Intrusion Detection System Based on a New Pattern Matching Algorithm" International Conference on Computer Engineering and Technology 2009 IEEE.

Author Biographies

K.V.S.N.Rama Rao was born in Andhra Pradesh, India and has attained his Masters in Computer Applications. from Andhra University and currently pursuing PhD in Berhampur University. His research areas are network security and Intrusion detection systems.

Pandu Prudvi has around 15 years of experience in software development field and currently working as SOA architect in Mahindra Satyam, Hyderabad. Has rich expertise in service oriented architecture and web services.

Manas Ranjan Patra was born in Orissa and has attained his Doctorate from Hyderabad Central University and currently working in Berhampur University. His research areas are network security, Intrusion detection systems and intelligent systems.

Performance Analysis of Parallel Mining for Association Rules on Heterogeneous System

Rakhi Garg[1], P.K.Mishra[2]
[1]Computer Science Section, MMV, Banaras Hindu University,
[2]Department of Computer Science, Banaras Hindu University,
Varanasi-221005, India
{rgarg, mishra}@bhu.ac.in

Abstract**:** Association Rule Mining plays an important role in predicting business trends those can occur in near future because it finds the hidden relationships among items in the transactions. Several sequential algorithms have been developed for finding maximal frequent itemsets and generating association rules. Due to advent of high storage devices large database can be stored. Parallel algorithms are very promising to mine these huge databases. Par-MaxClique, a parallel association rule mining algorithm is developed, uses static load balancing. In this paper we propose a simple parallel algorithm for association rule mining on heterogeneous system with dynamic load balancing based on Par-MaxClique algorithm. We compare our algorithm with the existing one for homogeneous environment and observed that the execution time gets reduced dramatically**.**

Keywords: Parallel association rule mining, heterogeneous system, Par-MaxClique algorithm

1. Introduction

Most of the parallel association rule mining algorithm developed so far uses static load balancing for homogeneous systems [12]. In static load balancing the job is initially partitioned among the homogeneous processors using some heuristics. There is no data movement among the processors during execution.

Moreover, if we apply the parallel algorithm developed for homogeneous system to heterogeneous environment, it will again leads to significant performance deterioration [1]. Since in homogeneous system there is an equal distribution of job among the processors of the same speed, uses static load balancing technique whereas heterogeneous system has processors of different speeds in which one completes job earlier than the other due to speed mismatch [4]. The high speed processor executes the assigned job quickly and sits idle while low speed processor is still busy with the assigned job that degrades the performance of the system. To utilize system processors efficiently and enhance the performance we design an algorithm that during execution checks the load of the processor and on the basis of which it moves the job from heavy loaded processor to least loaded one so that no processor sits idle till the completion of the whole jobs in a system.

In our algorithm, initially, the same number of jobs assigned to all the processors in a cluster by the scheduler using the same heuristics as in the homogeneous system. Since the processing speeds of the processors in the cluster are different so algorithm first finds out the fastest processor in the cluster and also computes the execution time to complete the execution of all the jobs assigned to it. After that it will compute the total number of complete and incomplete jobs of all the processors in the system and maintain load value of each processor in the cluster at the scheduler end i.e. the host. Then the load value of processors in a cluster are compared and the job is moved from the heavy loaded processor to the least loaded one and thus balances load dynamically in a cluster. A linked list containing the load values of all processors in a cluster are maintained at the scheduler end that gets updated during the completion of all the jobs assigned to the fastest processor. In this way load balancing becomes dynamic and involves data movement among the processor only when there is no communication overhead to enhance the performance of the system.

Section 2 briefly explains Par-MaxClique algorithm and focuses on the related work done. In section 3 we explain the functioning of the algorithm designed by us. Our experimental study is presented in section 4 and our conclusion in section 5.

2. Par-MaxClique Algorithm & Related Work

2.1 Par-MaxClique Algorithm

M. Zaki, Parathasarathy, Oghihara and Li [2] developed Par-MaxClique algorithm that gives more accurate frequent itemsets. It uses clique clustering which is more accurate than equivalence class clustering [2],[7]. Here, the database is vertically partitioned and hybrid search is applied on it to generate the longest frequent itemsets by using the (L_2) frequent 2-itemsets and some non frequent itemsets. The items are organized in a subset lattice search space, which is decomposed into small independent chunks or sub-lattices, which can be solved in memory. Efficient lattice traversal techniques are used, which quickly identify all the frequent itemsets via simple tid-list intersections [2].

Basically Par-MaxClique algorithm is divided into three phases i.e. initialization phase, asynchronous phase and final reduction phase [2],[7]. It generates clusters from L_2 using uniform hypergraph cliques and partition the clusters and the tid-list among the processors in the very first phase called the initialization phase. After that in the next phase called the asynchronous phase, the frequent itemsets are computed independently by each processors from the cliques assigned to it. Finally, the last phase i.e. the reduction phase produces the aggregate results and outputs the associations between the frequent itemsets.

EXAMPLE OF PAR-MAXCLIQUE ALGORITHM

Let database contains A,C,D,T and W four itemsets and 6 transactions are:-

Transactions	A	C	D	T	W
T1	1	1	0	1	1
T2	0	1	1	0	1
T3	1	1	0	1	1
T4	1	1	1	0	1
T5	1	1	1	1	1
T6	0	1	1	1	0

Tid-list is computed as: T(A) = {1,3,4,5}; T(C)={1,2,3,4,5,6}; T(D)={2,4,5,6} and T(W)={1,2,3,4,5}. During the initialization phase the tid-list is communicated among the processors and support counts for 2-itemsets are read. e.g. support count for AC ={1,3,4,5} = 4 which is counted by the intersection of the tid list of A and C. Similarly the support counts of AD, AT, AW, CD, CT, CW, DT, DW and TW are 2,3,4,3,4,4,3,2,3 and 3 respectively. Let us assume that minimum support = 3 so AD and DT will be discarded.

Frequent 2- itemsets are :-
AC,AT,AW,CD,CT,CW,DW,TW

Equivalence classes are:-
[A]: C T W
[C]: D T W
[D]: W
[T]: W

By applying the hypergraph clique for clustering to L2, the set of potential maximal cliques per equivalence class are generated.

Generated Maximal cliques per class:-
[A]: ACTW, ACW, ATW, ACT
[C]: CDW, CTW

Maximal cliques for equivalence class A

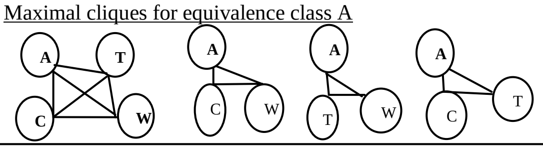


Figure 1: Equivalence class and Uniform Clique clustering [10]

Here, two cliques and equivalence class are generated which are distributed on the processors to achieve equal load balancing. Each processor independently computes the maximal frequent itemsets which are used in association rule generation in the last phase of the algorithm.

Par-MaxClique algorithm uses the static load balancing technique with some heuristics for equal balance among the processors in the homogeneous system. This is far from reality because a database server has multiple systems with different configurations and speeds. If this algorithm is used there then it will degrade the performance of the system. This demands the dynamic load balancing schemes.

We have developed an algorithm for parallel mining of the association rules for such heterogeneous system that uses dynamic load balancing technique and enhances the system performance by reducing the execution time.

2.2 Related Work

Several parallel algorithms for association rules have been proposed in the literature. The most known parallel algorithm is Count Distribution (CD), Data distribution and Candidate Distribution; proposed by Rakesh Agrawal and J. Shafer [4], [5], [6]. Among these CD is the most promising one which minimizes the communication overheads but utilize memory less efficiently than DD.

The FDM and FPM algorithms are the enhanced versions of CD [3]. In FDM, two rounds of the communications are required in each iteration one for computing the global support and the other for broadcasting the frequent itemsets. FPM is more efficient than FDM in communication which broadcast local supports to all processors which is determined by the candidate size as in CD [3]. Thus for small minimum support, the communication cost could be very high at some passes where the candidate set is large.

Par-MaxEclat and Par-MaxClique is parallel MFI (maximal frequent itemsets) mining algorithm proposed which are parallel versions of MaxEclat and MaxClique respectively. These algorithms distribute over the processor in the system the cluster of the generated potential maximal frequent itemsets. These algorithms are implemented on dedicated homogeneous system which uses static load balancing technique. Par-MaxClique algorithm outperforms CD algorithm because it utilizes the aggregate memory of the parallel system, decouples the processors right in the beginning by repartitioning the database so that each processor can compute independently, use vertical database layout which clusters the transactions containing an itemset into tid-list without scanning the database and computes the frequent itemsets by simple intersections on two tid-lists without having an overhead of maintaining complex data structures[2].

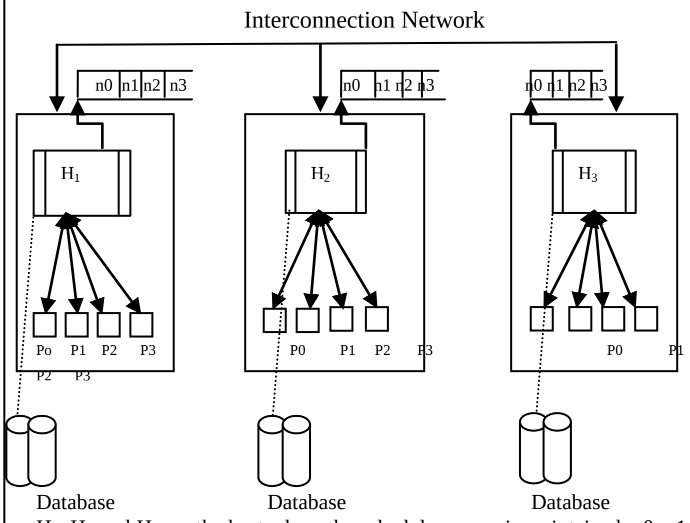
Problem here is that although Par-MaxClique algorithm outperforms but it has limitation that it is only implemented for homogeneous system that uses static load balancing technique. It won't take care of fault tolerance i.e. what happens if one of the processor in the system fails, how the jobs assigned to it gets executed and also what happens in the case of heterogeneous system which have processors with speed mismatch. If it used in heterogeneous system with no check on the load factor of the processor maintained during execution phase then it might happen the processors with high speed may sit idle after completing the execution of all the jobs assigned to it while others with less speed are still involved in the processing work. This won't utilize the processor to their maximum extent. Hence an algorithm which uses dynamic load balancing technique is needed for proper utilization of all the processors in the cluster.

Load Balancing FP-Tree (LFP-tree) algorithm is proposed by Kun-Ming Yu, Jiayi Zhou and Wei Chen Hsiao based on FP-tree structure that divides the item set for mining by evaluating the tree's width and depth and proposed a simple and trusty calculate formulation for loading degree [8]. But it has limitation of maintaining the complex tree structure.

Masaru Kitsuregawa and Takahilus Shintani, Masahisa Tamura and Iko Pramudiono, proposed Parallel Data Mining on large scale PC Cluster, the dynamic load balancing methods for association rule mining for heterogeneous system [9] which uses candidate migration and transaction migration. Initially if load is not balanced after candidate migration then it applies the transaction migration which is costly but more effective for strong imbalance.

3. Proposed Algorithm

In our algorithm, the hypercliques of frequent 2-itemsets which are considered as jobs are equally divided among the processors of the system for having equal load balance as done in the homogeneous system. During execution we find out the processor which has completed the execution of all the jobs assigned to it i.e. the fastest processor of the cluster. Then we arrange the processors in the decreasing order according to their respective speeds and compute the number of complete and incomplete jobs of every processor at a time when fastest processor have completed the execution of all jobs assigned to it. After that the scheduler queue which is maintained at the host to which numbers of processors are attached and contains the load value i.e. the number of incomplete jobs of every attached processor gets updated. After that the data is moved from heavy loaded i.e. the slowest processor to the least loaded one i.e. the fastest processor in the cluster only if the remaining execution time of the job assigned to the slowest processor is more than that of its execution time at the fastest processor. This takes care of communication overhead. Since we have distributed the hypercliques among the processors in the cluster for generating the maximal frequent itemsets (MFI), it might happen that the fastest processor have generated it at a time when others are involved in generating MFI from one of the cliques from the cluster of cliques assigned to it. In that case the remaining untouched cliques from the list of a given processor will move to the fastest processor for computation. This will engage all the processors of various speeds in the cluster which cannot be done by adopting the algorithm designed for the homogeneous system. In this way every processor in the cluster gets utilized to its maximum extent and also reduces the total execution time. e.g. Consider a case where (frequent 2-itemsets) L_2 = {12, 13, 14, 15, 23, 24, 25, 34, 35, 45} and two Processors P_o and P_1; where P_o is faster than P_1. For having equal load balance the clique of [1] get assigned to P_0 while [2] and [3] get assigned to P_1. It might happen that P_0 have generated all the MFI from clique [1] at a time when P_1 is busy in generating from [2] and [3] remained untouched. In that case [3] gets moved to P_0.



H_1, H_2 and H_3 are the host where the scheduler queue is maintained. n0, n1, n2 and n3 are the load factor of P_0, P_1, P_2 and P_3. Here, the database of tid-list is initially equally partitioned and the support count of all L_2 are available.
Each processor say P_0, P_1, P_2 and P_3 computes the maximal frequent itemsets from the cluster assigned to them.

Figure 2: Working of algorithm in Heterogeneous system

Table 1: Pseudo code for parallel association rule mining algorithm for heterogeneous environment

```
Begin
/* Initialization Phase */
1. Generate $L_2$ from 2-itemset support counts
2. Generate clusters from $L_2$ using uniform hypergraph cliques
3. Partition clusters among the processors
4. Scan local database partition
5. Transmit relevant tid-list to other processors
6. Received tid-list from other processors.
7. First, we compute the job queue and linked list of each processor and scheduler respectively. Initially, all processor have the same load value since jobs are equally distributed among the processors as in Par-MaxClique algorithm for homogeneous system.

/* Asynchronous Phase */
8. For each assigned cluster $C_2$, compute Frequent Itemsets
9. During execution, each processor updates its job queue and the linked list at the scheduler is also gets updated accordingly.

/* Communication OR Complete and offer Phase */
10. If job queue of all processors are empty then stop
11. else
The scheduler compares the load value of all the processors within the cluster and if any difference is found then perform the following :-
(i) Job from heavy loaded processor say $P_i$ is taken and gets assigned to least loaded processor say $P_j$.
(ii) Job queues of the $P_i$ and $P_j$ are adjusted accordingly.
(iii) The link list at the scheduler is also adjusted accordingly.
12. Go to asynchronous phase i.e. step 8.

/* processing completes at each processor and then moves to reduction phase that involves 13.*/
13. Aggregate Results and Output Associations
14. STOP
```

4. Analysis of proposed algorithm

We have designed a simulator in C language that reads number of processor in the system, there processing speed and the number of jobs to be executed by the system. The execution time of each of the job is randomly generated. The major difference between the homogeneous and heterogeneous system is observed at the Communication OR Complete and offer Phase of the proposed algorithm where dynamic allocation of jobs are done in heterogeneous system and static in case of homogenous system. Initially our simulator distributes the jobs equally among all the processors in the system so that work load remains same at every processor and then computes the actual execution time of each processor as well as the computation time of all the jobs assigned to it for doing dynamic allocation. In the case of heterogeneous system the actual execution time of each processor is different whereas it remains same in the case of homogeneous system. So, simulator will list out the total number of incomplete jobs allocated to each processor at the time when the fastest processor has completed the execution of all the jobs assigned to it. On the basis of that the entry at the scheduler that keeps track of the work load factor of each of the processor in the system will be updated. After that simulator compares the remaining execution time of the incomplete jobs of each processor with its execution time at the fastest processor and if it is more then only the data movement will be done from that processor to the fastest processor otherwise not. In this way the fastest processor is not overloaded and this process repeats till all the jobs complete its execution. By doing so it will also take care of fault tolerance because if any of the processor is not completing its execution then after comparing it with the processor arranged in the decreasing order of their processing speed the job will be assigned to the one that involves in processing. Not only this, it also does the equal distribution of jobs among the processors in the system while doing dynamic allocation so that no processor sits idle. We can observe it very well in figure 7.

We have executed algorithm in heterogeneous system having four processors with processing speeds 2.2GHz, 3.2GHz, 3.6GHz and 3.8GHz for the number of jobs executed ranging between 200 and 15,000 and also the same in homogeneous system with four processors with processing speed 2.2GHz, 3.2GHz, 3.6GHz and 3.8GHz respectively and obtain results shown in figure 3,4,5 and 6.

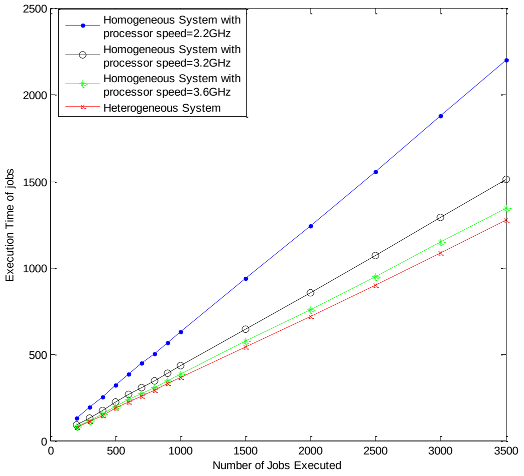


Figure 3

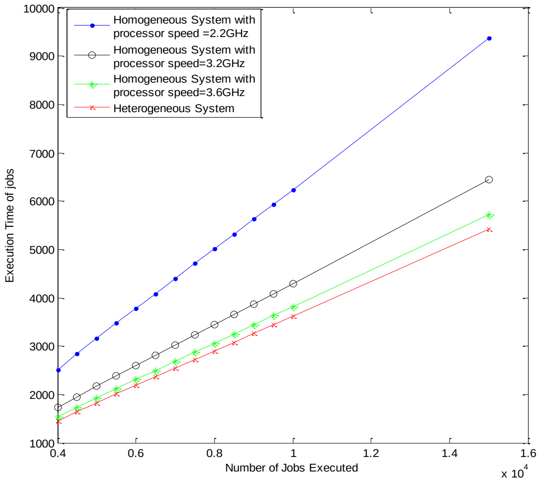


Figure 4

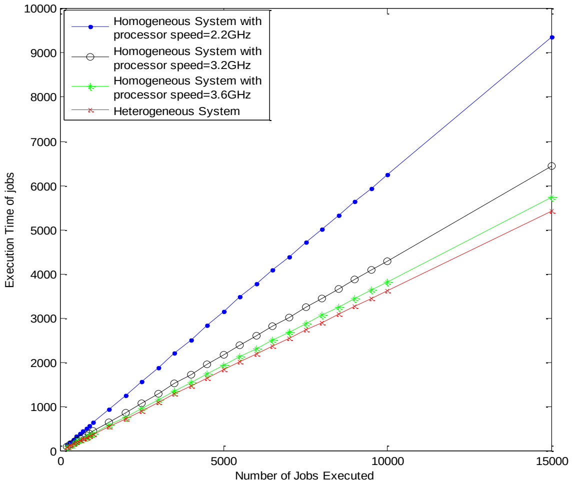


Figure 5

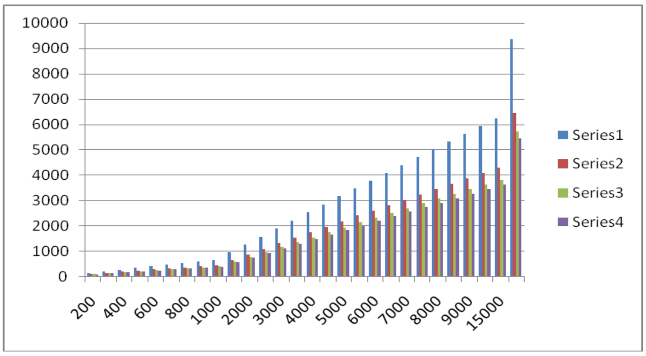


Figure 6

In figure 6, Series1, Series2 and Series3 represents homogeneous system having 4 processors with processing speed 2.2GHz, 3.2GHz and 3.8GHz respectively, Series 4 showing heterogeous system having 4 processors with processing speed 2.2GHz, 3.2GHz, 3.6GHz and 3.8GHz.

It is observed from figure 2 and 3 that the execution time reduces dramatically when the number of jobs increases above 10000 as compared to it range between 200 and 4000. It means that as the number of jobs increases the execution time reduces. Thus, we can say that the performance of the heterogeneous system which uses dynamic load balancing is much better than that of homogeneous system that uses static load balancing technique. Moreover, we can save cost also by having some low speed processors because instead of having all high speed processors, the same performance can be achieved by having a combination of low and high speed processors in the cluster. Not only is this it also seen that the high performance is obtained in the heterogeneous system as compared to the homogeneous with the involvement of the same number of processors in the data movement which is shown in figure 7.

5. Conclusion

In this algorithm, we have reduced the execution time. Moreover, the same performance can be achieved by the heterogeneous system with a combination of low and high speed processors as in homogeneous system with same number of high speed processors. It means we can utilize the low speed processors also for having the desired performance. Hence the cost of having all high speed processors can be saved. Not only this, our algorithm also takes care of fault tolerance in the cluster. Also, it has good features of the Par-MaxClique parallel association rule mining algorithm for homogeneous environment which outperforms count distribution, data distribution and candidate distribution algorithm for parallel association rule mining. It enhances the performance of the heterogeneous system by having dynamic load balancing techniques.

In future, we try to perform dynamic load balancing in between the clusters. If the graph is too dense and if support decreases and transaction size increases it will affect the edge density and leads dense graph resulting in large cliques with significant overlap among them. We will try to handle this problem in our future work.

6. References

[1] Masahisa Tamura and Masaru Kitsuregawa, “Dynamic Load Balancing for Parallel Association Rule Mining on Heterogeneous PC Cluster System”, Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, pp. 163, 1999.

[2] Mohammed J. Zaki, Srinivasan Parthasarthy, Mithsunori Ogihara and Wei Li, “Parallel Algorithms for Discovery of Association Rules”, Data Mining and knowledge Discovery, © Kluwer Academic Publishers, pp. 360, 364, 1997.

[3] Soon M. Chung, Congnan Luo, “Efficient mining of maximal frequent itemsets from databases on a cluster of workstations”, © Springer-Verlag London Limited 2007 pp. 359-391, Published online: 12 December 2007.

[4] M. J. Zaki, “Parallel and Distributed Association Mining: A Survey”, Concurrency, IEEE, Volume 7, Issue 4, pp. 2-3, 10-13, Oct-Dec. 1999.

[5] R. Agrawal and J. Shafer, “Parallel mining association rules”, IEEE Transaction On Knowledge and Data Engineering, Volume 8, Issue 6, 8(6): pp. 962-969, December 1996.

[6] E-H. Han, G. Karypis and Vipin Kumar, “Scalable parallel data mining for association rules”, Proceedings of ACM SIGMOD Conf. Management of Data, pp. 279-284, May 1997.

[7] Mohammed J. Zaki, “Parallel and Distributed Data Mining: An Introduction”, C.-T. Ho (Eds.): Large-Scale Parallel Data Mining © Springer-Verlag Berlin Heidelberg, LNAI 1759, pp. 9, 2000.

[8] Kun-Ming Yu, Jiayi Zhou and Wei Chen Hsiao, “Load Balancing Approach Parallel Algorithm for Frequent Pattern Mining”, V. Malyshkin (Ed.): PaCT 2007. © Springer-Verlag Berlin Heidelberg. LNCS 4671, pp. 623–631, 2007.

[9] Masaru Kitsuregawa and Takahilus Shintani, Masahisa Tamura and Iko Pramudiono, “Parallel Data Mining on large scale PC Cluster”, H. Lu and A. Zhou (Eds.): WAIM 2000, © Springer-Verlag Berlin Heidelberg, LNCS 1846, pp. 15–26, 2000.

[10] Rakhi Garg and P. K. Mishra ,“Parallel Association Rule Mining on Heterogeneous system”, research papers published in an International Journal of Computer Application (0975 – 8887) , Volume-1, No. -14, pp. 87-91; Feb 2010.

[11] Jochen Hipp, Ulrich Gauntzer, Gholamreza Nakhaeizadeh,”Algorithms for Association Rule Mining-A General Survey and Comparison”, SIGKDD explorations copyright © 2000, ACM SIGKDD, Volume 2, Issue 1, pp. 58-61, July 2000.

Comparative Genome Analysis using Computational Approach in HIV-1 and HIV-2

Dr. DSVGK Kaladhar[1] and A.Krishna Chaitanya[2]

[1,2]Department of Bioinformatics, GIS, GITAM University, Visakhapatnam. Andhra Pradesh,

Corresponding Address
dr.dowluru@gmail.com

***Abstract*:** Comparative genome analysis are playing important role in genomics, where the scientific studies can provide the relationship of two or multiple organisms. Both HIV-1 and HIV-2 predicted nine genes in entire genome. HIV is a nano particle (sometimes living organism) has molecular information which can use host machinery in the formation of proteins such as gag polyprotein, gag-pol polyprotein, vif, vpr, tat, vpu, gp160, envelope glycoprotein and Nef located in various regions of the genome. Vpu is not predicted in HIV-2. The three dimentional molecules can further be used for the identification of drug targets for the control of diseased molecules located in HIV.

***Keywords*: HIV-1, HIV-2, gene prediction, modelling.**

1. Introduction

Human immunodeficiency virus (HIV) is a retrovirus [1] that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system [2], [3] begins to fail, leading to life-threatening opportunistic infections.

The human brain often compared to digital computer [4], [5]. There are various logistic methods to explore the nature of life. Earth is a system which contains power of life. Humans are also be compared to earth with complex system dominating other systems [6].Viruses though have simple system, it has the power in control of complex systems such as plants and animals, and even humans. HIV is one of the powerful machine (have both living and non-living features) [7] have capability to control human cellular and immune systems. The relationships between organisms, such as those between prey or predator, host and parasite, and between mating partners, are complex and multidimensional [8].

Bioinformatics is a science increasingly essential to navigate and manage the host of information generated in cellular systems: to improve study design, make candidate gene identification, interpret and manage data, and to explore light on the molecular pathology of disease-causing mutations [9]. In the genome age, after completion of Human genome project (HGP), a major research goal is to find the functions of genes and to define their interactions in a particular organism [10].

Several disciplines such as genomics, proteomics, immunoinformatics and systems biology have recently emerged within bioinformatics which for the first time enable biological systems to be studied on a scale commensurate with there inherent complexity. Most of these studies are consequently assuming a central play in modern drug development, with a wide spectrum of practical applications embracing target discovery, target validation, lead compound selection, investigation of drug modes of action, diagnostics, toxicology and clinical development [11].

In 1983, scientists led by Luc Montagnier at the Pasteur Institute in France first discovered the virus that causes AIDS [12]. The structure of HIV is different from other retroviruses and is about 120 nm in diameter (120 billionths of a meter; around 60 times smaller than a red blood cell) and roughly spherical [13].

HIV-1 and HIV-2 are two species of HIV infect humans. HIV-1 is thought to have originated in southern Cameroon after evolving from wild chimpanzees (*Pan troglodytes*) to humans during the twentieth century. HIV-2 is largely confined to West Africa [14].

2. Methodology

Biological databases creates public databases which are conducting research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information for the better understanding of molecular processes affecting human health and disease. A complete genome of Human immunodeficiency virus 1 ACCESSION NC_001802 (9181bp) and Human immunodeficiency virus 2 ACCESSION NC_001722 (10359 bp) were selected for the present study.

FGENESV0

FGENESV0 is the fastest and most accurate *ab initio* gene prediction program for viruses. Its variants that use similarity information in gene prediction, resulting in fully automatic annotation of quality similar to that of manual annotation.

Protein Molecular Weight

Protein Molecular Weight accepts a protein sequence and calculates the molecular weight of the submitted protein sequence. Protein molecular weight is calculated using tolls provided from ExPASy server.

Protein Isoelectric Point

Protein Isoelectric Point calculates the theoretical pI for the submitted protein sequence. Isoelectric point of a protein is calculated using ExPASy server.

BLASTP
BLASTP accepts protein (AA) sequences and compares them against the protein databases. The BLAST (Basic Local Alignment Search Tool) programs have been designed for speed to find high scoring local alignments. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity. Because of its design for speed, there may be a minimal loss of sensitivity to distant sequence relationships.

SWISS-MODEL
SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server. The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists World Wide.

3. Results

HIV is a single stranded, linear, RNA containing organism. Nine genes are predicted in HIV-1 by viral gene prediction server and the genes characterized by BLASTP shown as:
Gene 1 Gag Polyprotein
Gene 2 Gag-Polpolyprotein
Gene 3 vif (viral infectivity factor)
Gene 4 Protein Vpr (Viral protein R)
Gene 5 tat protein(p28-tev)
Gene 6 Protein Vpu (Viral protein U)
Gene 7 Envelope surface glycoprotein gp160
Gene 8 envelope glycoprotein
Gene 9 Protein Nef(Negative factor)(F-protein)

In HIV-2 genome also predicted nine genes by FGENEV0 server and are shown as:
Gene 1 nef gene
gene 2 gag-pol fusion protein
gene 3 gag polyprotein
gene 4 gag pol fusion protein
gene 5 vif protein
gene 6 vpr protein
gene 7 tat protein
gene 8 env polyprotein
gene 9 nef protein .

Gag-Pol Polyprotein is having highest molecular weight and tat protein is having lowest molecular weights in both HIV-1 and HIV-2. All the proteins stands between 4 to 11 pH levels which has shown that the proteins of viruses can denature or become inactive, if there is a decrease of salinity below 4 pH or above 11 pH (Table 1 and 2).

The identification and characterization Viral protein U is absent in HIV-2 and is present in HIV-1. This protein may provide highest infectivity to humans and can have the capability in the control of human immune system.

Table 1: HIV-1 gene identification and characterization

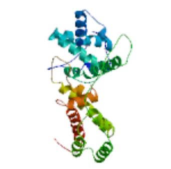
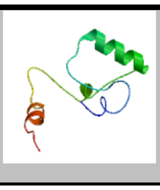
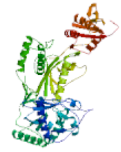
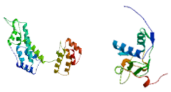
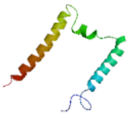
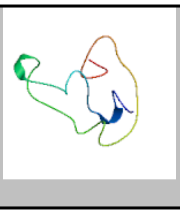
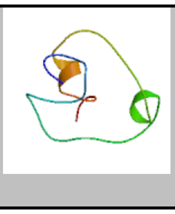
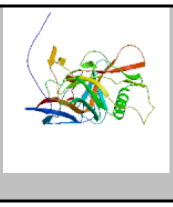
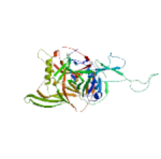
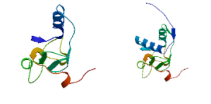
Gene No.	**Type**	**Mol.wt kDs**	**pI**
Gene1	Gag polyprotein (Pr55Gag)	55.94	9.42
gene2	Gag-Pol polyprotein (Pr160Gag-Pol)	103.58	8.93
gene3	vif (viral infectivity factor)	22.52	10.39
Gene4	Protein Vpr (Viral protein R)	9.31	4.76
Gene5	tat protein	8.39	10.48
Gene6	Protein Vpu (Viral protein U)	9.24	4.43
Gene7	Envelope surface glycoprotein gp160	97.23	9.19
Gene8	envelope glycoprotein	20.24	7.12
Gene9	Protein Nef (Negative factor) (F-protein)	13.69	6.96

Table2: HIV-2 gene identification and characterization

Gene No.	**Type**	**Mol.wt kDs**	**pI**
Gene1	Nef (Negative factor)	15.03	5.57
gene2	Gag-Pol polyprotein (Pr160Gag)	45.89	9.56
gene3	Gag polyprotein (Pr55Gag)	11.50	6.80
Gene4	Gag-Pol polyprotein (Pr160Gag	110.09	8.95
Gene5	Virion infectivity factor (Vif)	25.32	10.18
Gene6	Protein Vpr (Viral protein R)	10.07	6.96
Gene7	tat protein	11.24	9.65
Gene8	Envelope glycoprotein gp160	98.95	8.79
Gene9	Protein Nef (Negative factor)	29.92	6.11

Modelled structures of these nine proteins are provided in Table 3.

Table 3: Protein modelling using Swissmodel

Protein	Protein modelling HIV-1	Protein modelling HIV-2
Gag polyprotein (Pr55Gag)		
Gag-Pol polyprotein (Pr160Gag-Pol)		
vif (viral infectivity factor)	No Templates found	No Templates found
Protein Vpr (Viral protein R)		
tat protein		
Protein Vpu (Viral protein U)		Gene not predicted
Envelope surface glycoprotein gp160		
envelope glycoprotein	No Templates found	Gene not predicted
Protein Nef(Negative factor)(F-protein)		

4. Discussion

Information Technology and Biological sciences are being transformed due to enormous growth of data from laboratories worldwide. Most of the biologists and computer scientists focus to explore innovations of their research in faster rate using developments in Information technology. A complete genome of HPV-92 predicts six genes by gene prediction technique [15]. The comparison of two or more sequences of numbers or letters is common in several fields such as molecular biology, bioinformatics, speech recognition and computer science [16].

Hence by alignment of strings in the genome of organisms can predict the nature and functions of various cellular systems, which can in turn predict the health of the species. A small organism such as HIV has the capacity to kill the complex organisms such as Humans. Hence the scientists are focusing on how the nanoorganisms such as viruses are destroying the complex organisms which are measuring in few feet's of high. There are compounds which are still lesser in sizes than nano and have the capabilities to change the cellular processes of life.

The present studies have provided the string which can have the capacity to process the data. The stored information in HIV can lead the process of 3D compounds and has the capability to control the cellular mechanisms of human life. The studies of data provided the sizes and pI values which are the biological properties of HIV. The nine proteins of HIV-1 and HIV-2 can be controlled, if all the systems mechanisms are predicted using the advanced information technologies.

5. Conclusion

Predictions of molecular structures are highly necessary in the studies of systems biology. The increased data availability can provide answers for the life process from birth to death. Hence the high end processing of data can explore the characterization and functional cellular changes in various species which are measuring from nano to larger sizes.

6. Acknowledgements

The authors would like to thank GITAM University for providing computational facility and access to e-journals to carry out this research.

References

[1] J. F. Roeth, K. L. Collins, "Human Immunodeficiency Virus Type 1 Nef: Adapting to Intracellular Trafficking Pathways", Microbiol. Mol. Biol. Rev., Vol 70, 548-563, 2006.

[2] J. B. Alimonti, T. B. Ball, K. R. Fowke, Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS, J. Gen. Virol., Vol. 84, 1649-1661, 2003.

[3] AS Fauci, "The human immunodeficiency virus: infectivity and mechanisms of pathogenesis",Science, Vol. 239, Issue 4840, 617-622, 1988.

[4] Collins, D.L.; Zijdenbos, A.P.; Kollokian, V.; Sled, J.G.; Kabani, N.J.; Holmes, C.J.; Evans, A.C., "Design and construction of a realistic digital brain phantom ", Medical Imaging, V 17, No.3,pp. 463 – 468, 1998.

[5] Hans Moravec, "When will computer hardware match the human brain?",Journal of Evolution and Technology. 1998. Vol. 1, 1998.

[6] Peter M. Vitousek, Harold A. Mooney, Jane Lubchenco, Jerry M. Melillo, Human Domination of Earth's Ecosystems“,Science, Vol. 277. no. 5325, pp. 494 – 499, 1997.

[7] Ralph S. Lillie, “Living Systems and Non-Living Systems”, Philosophy of Science, Vol. 9, No. 4, pp. 307-322, 1942

[8] Vladimir B. Bajić, Tin Wee Tan, “Information Processing and Living Systems”, Imperial College Press, pp. 1, 2005.

[9] Michael R. Barnes, Ian C. Gray, “Bioinformatics for Geneticists”, John Wiley & Sons, pp. 6, 2003.

[10] David W. Mount, “Bioinformatics: Sequence and Genome Analysis”, Cold Spring Harbor Laboratory Press, pp. 3, 2004.

[11] Alexander Hillisch and R. Hilgenfeld, “Modern methods of drug discovery”, Birkhäuser Basel, 1 edition, pp. 19-20, 2003.

[12] Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)", Science, vol. 220, no. 4599, pp. 868-871, 1983.

[13] McGovern SL, Caselli E, Grigorieff N, Shoichet BK, "A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening", J Med Chem, vol. 45, no. 8, pp. 1712-1722, 2002.

[14] Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H., "Origin of HIV-1 in the Chimpanzee Pan troglodytes troglodytes", Nature, vol. 397, no. 6718, pp. 436-441, 1999.

[15] D.S.V.G.K.Kaladhar, T. Uma Devi, P.V. Nageswara Rao, An *in silico* genome wide identification, characterization and modeling of Human Papilloma Virus strain 92**,** IJEST, Vol. 2, no. 9, pp. 4288-4291, 2010.

[16] Satyasaivani B., Kaladhar DSVGK, Shashi M. and Kesavareddy J., “Prediction of *methanobacterium* using suffixtree”, International Journal of Machine Intelligence, Vol. 1, no. 2, pp-1-4, 2009.

Author Biographies

Dr.DSVGK Kaladhar: Dr. Kaladhar was born in Vijayawada on 1974. He has got his doctoral degree in Biotechnology and M,Sc in Microbiology from ANU, Guntur. He is presently working as Asst. Professor in Bioinformatics, GITAM University, Visakhapatnam. His field of study is neural networks, Image analysis, Bioinformatics and Biology

Mr. A.Krishna Chaitanya: Mr. Chaitanya was born in Rajahmundry on 1984. He has acquired Master of Science in Biochemistry from AU, Visakhapatnam. He is presently working as Asst.Professor in GITAM University, Visakhapatnam. His specializations are Molecular modeling, Bioinformatics and Biochemistry.

A Study of Neutrosophic technology to retrieve Queries in Relational Database

Ashit Kumar Dutta [a]*, Ranjit Biswas [b], Nasser Saad AL-Arifi [c]

[a] Department of Computer Science and Information System
Faculty of Information Technology
SHAQRA University ,Shaqra , Saudi Arabia
drashitkumar@yahoo.com

[b] Department of Computer Science and Information Technology
ITM University, Gurgaon ,Haryana ,India
ranjitbiswas@yahoo.com

[c] Department of Geophysics, College of Science
King Saud University, Riyadh, Saudi Arabia
nalarifi@ksu.edu.sa

Abstract: *In this study the authors propose a new method of searching techniques called Neutrosophic-search to find the most suitable match for the predicates to answer any imprecise query made by the database users. It is also to be mentioned that the Neutrosophic-search method could be easily incorporated in the existing commercial query languages of DBMS to serve the lay users better. So in this study Authors are suggesting a new method called as α-Neutrosophic-equality Search to answer the imprecise queries of Relational database based on ranks.*

Keywords: *Neutrosophic logic, α-neutrosophic-equality search, neutrosophic search, β-value of an interval, probabilistic database.*

1. Introduction

Today Databases are Deterministic. An item belongs to the database is a probabilistic event, or a tuple is an answer to the query is a probabilistic event and it can be extended to all data models. Here we will discuss probabilistic relational data. Probabilistic relational Data are defined in two ways, Database is deterministic and Query answers are probabilistic or Database is probabilistic and Query answers are probabilistic.

Probabilistic relational databases have been studied from the late 80's until today. But today Application Need to manage imprecision's in data. Imprecision can be of many types: non-matching data values, imprecise queries, inconsistent data, misaligned schemas, etc.

The quest to manage imprecision's is equal to major driving force in the database community is the Ultimate cause for many research areas: data mining, semi structured data, and schema matching, nearest neighbor. Processing probabilistic data is fundamentally more complex than other data models.. Now our implementation includes Ranking query answers. Since our Database is deterministic, The query returns a ranked list of tuples But our User interested in top-k answers. Sometimes we get the empty answers for the user queries in the deterministic database. For e.g.,

For example, consider a database of personal Computers,

```
Select * from PC
Where cpu = '8086'
And memory = 8
Rank_by clock_rate >= 25
                    Disk_size >= high
               access_time <25
                     price = low
```

Here our Database will fail to Answer because of the imprecision in the query. But Using Ranking query using the neutrosophic logic we will get the answer. So to Answer this we must know the type of imprecision.

Definition Ranking: Ranking is defined as Computing a similarity score between a tuple and the query, Consider the query

```
Q = SELECT*
      From R
      Where A1= v1 and … and Am = vm
```
Query is a vector: Q = (v1,…, vm)
Tuple is a vector: T = (u1,…, um)

Consider the applications: personalized search engines, shopping agents, logical user profiles, soft catalogs.

To answer the queries related with the above application two approaches are given:

- Qualitative$\rightarrow$Pare to semantics (deterministic)
- Quantitative $\rightarrow$ alter the query ranking

Definition: An imprecise attribute value t_m (a_i) must be specified as a discrete probability distribution over D_i, that is $t_m(a_i) = \{(z_j,P_j)\backslash z_j \in D$iand $P_j \in [0, 1]\}$ with $\Sigma P_j = \alpha_{im}$, $0< = \alpha_{im} < = 1$. $(z_j,P_j) \in f_{vn}(a)$.

This definition covers both interpretations of null values as well as the usual interpretation of imprecise data: If $a_{im} = 1$, we certainly know that an attribute value exists and with $a_{im} = 0$, we represent the fact that no value exists for this attribute. In the case of $0<o_i$, <1, o_i, gives the probability that an attribute value exists: For example, someone who is going to have a telephone soon gave us his number, but we are not sure if this number is valid already. With imprecise values specified this way, their probabilistic indexing weight can be derived easily.

Definition probabilistic tuples: Let R (A) be a relation scheme and let $t = (V_1; : : : ; V_n)$ be a tuple of cases of the relation scheme R. For each V_i, let V_I be the set of the $v_j = (a_j, l_j, u_j; p_j)$ such that $(a_j ; l_j ; u_j) \in V_i$, where p_j is the path associated with aj . A probabilistic tuple $t_0 = (v_1'; : : : ; v_n')$ is an element of the Cartesian product $V_1\times \ldots\ldots \times V_0$. By A_i. l, A_i. u and $A_{i.}$ p we denote l_j, u_j and p_j associated with a generic value of A_i in a given probabilistic tuple, respectively.

Definition: Probabilistic relation: A probabilistic relation r of the scheme R (A) is a finite set of probabilistic tuples of R (A). By domr (A_i) we will denote the set of all values of the attribute A_i in the relation r.

Definition: Probabilistic database: A probabilistic database of the database scheme R $= \{R_1(A_1), : : : ;R_m (A_m)\}$ is a finite set of probabilistic relations $r = (r1, \ldots\ldots, r_m)$, where each ri is a relation of the scheme $R_i(A_i)$. In order to avoid probabilistic ambiguities we assume that in each initial relation there cannot be identical tuples.

So the failure of the RDBMS due to the presence of imprecise constraints in the query predicate which cannot be tackled due to the limitation of the grammar in standard query languages which work on crisp environment only. But this type of queries is very common in business world and in fact more frequent than grammatical-queries, because the users are not always expected to have knowledge of DBMS and the query languages.

Consequently, there is a genuine necessity for the different large size organizations, especially for the industries, companies having worldwide business, to develop such a system which should be able to answer the users queries posed in natural language, irrespective of the query languages and their grammar, without giving much botheration to the users. Most of these type of queries are not crisp in nature and involve predicates with fuzzy (or rather vague) data, fuzzy/vague hedges (with concentration or dilation). Thus, this type of queries is not strictly confined within the domains always. This corresponding predicates are not hard as in crisp predicates. Some predicates are soft because of vague/fuzzy nature and thus to answer a query a hard match is not always found from the databases by search, although the query is nice and very real and should not be ignored or replaced according to the business policy of the industry. To deal with uncertainties in searching match for such queries, fuzzy logic and rather vague logic [1] and Neutrosophic logic by Smarandache[7] will be the appropriate tool.

In this study we propose a new type of searching techniques called as neutrosophic search which is a combination of α_ Neutrosophic-equality search and neutrosophic proximity search by using Neutrosophic set theory to meet the predicates posed in natural language in order to answer imprecise queries of the users. Thus it is a kind of an intelligent search for match in order to answer imprecise queries of the lay users. We call this method by Neutrosophic search which is a combination of α-Neutrosophic-equality search and neutrosophic proximity search.

Our method, being an intelligent soft-computing method, will support the users to make and find the answers to their queries without iteratively refining them by trial and error which is really boring and sometimes it seriously effects the interest (mission and vision) of the organization, be it an industry, or a company or a hospital or a private academic institution etc. to list a few only out of many. Very often the innocent (having a lack of DBMS knowledge) users go on refining their queries in order to get an answer. The users are from different corner of the academic world or business world or any busy world. For databases to support imprecise queries, our intelligent system will produce answers that closely match the queries constraints, if does not exactly. This important issue of closeness cannot be addressed with the crisp mathematics. That is why we have used the Neutrosophic tools.

Theory of neutrosophic set: In the real world there are vaguely specified data values in many applications, such as sensor information, Robotics etc. Fuzzy set theory has been proposed to handle such vagueness by generalizing the notion of membership in a set. Essentially, in a Fuzzy Set (FS) each element is associated with a point-value selected from the unit interval [0,1], which is termed the grade of membership in the set. A Vague Set (VS), as well as an Intuitionistic Fuzzy Set (IFS), are a further generalization of an FS.

Now take an example, when we ask the opinion of an expert about certain statement, he or she may say that the possibility that the statement is true is between 0.6 and 0.8 and the statement is false is between 0.3 and 0.5 and the degree that he or she is not sure is between 0.2 and 0.4. Here is another example, suppose there are 10 voters during a voting process. In time t1, two vote yes, three vote no and five are undecided, using neutrosophic notation, it can be expressed as x (0.2,0.5,0.3); in time t2, three vote yes, two vote no, two give up and three are undecided, it then can be expressed as x (0.3,0.3,0.2). That is beyond the scope of the intuitionistic fuzzy set. So, the notion of neutrosophic set is more general and overcomes the aforementioned issues. In neutrosophic set, indeterminacy is quantified explicitly and truth membership, indeterminacy-membership and falsity membership are independent. This assumption is very important in many applications such as information fusion in which we try to combine the data from different sensors. Neutrosophy was introduced by Smarandache[7].

Neutrosophic set is a powerful general formal framework which generalizes the concept of the classic set, fuzzy set[2], Vague set[1] etc.

A neutrosophic set A defined on universe U. x = x (T, I, F) ε A with T,I and F being the real standard or non-standard subsets of]0- ,1+[, T is the degree of truth membership of A, I is the degree of indeterminacy membership of A and F is the degree of falsity membership of A.

Definition: A Neutrosophic set A of a set U with $t_A(u)$, $f_A(u)$ and $I_A(u)$, $\forall u \in U$ is called the α-Neutrosophic set of U, where $\alpha \in [0,1]$.

Definition: A Neutrosophic number (NN) is a Neutrosophic set of the set R of real numbers.

Operations with neutrosophic sets: We need to present these set operations in order to be able to introduce the neutrosophic connectors.

Let S1 and S2 be two (unidimensional) real standard or non-standard subsets, then one defines.

Addition of sets: S1 ⊕ S2 = {x|x = s1+s2, where s1∈S1 and s2∈S2}, with inf S1⊕S2 = inf S1+inf S2, sup S1⊕S2 = sup S1+sup S2 and as some particular cases, we have {a}⊕S2 = {x|x = a+s2, where s2∈S2} with inf {a}⊕S2 = a+inf S2, sup {a}⊕S2 = a+sup S2.

Subtraction of sets: S1ϴS2 = {x|x=s1-s2, where s1∈S1 and s2∈S2}.

For real positive subsets (most of the cases will fall in this range) one gets inf S1ϴS2 = inf S1-sup S2, sup S1ϴS2 = sup S1-inf S2 and as some particular cases, we have {a}ϴS2 = {x|x = a-s2, where s2∈S2}, with inf {a} ϴS2 = a-sup S2, sup {a} ϴS2 = a-inf S2.

Multiplication of sets: S1⊕S2 = {x|x = s1.s2, where s1∈S1 and s2∈S2}.
For real positive subsets (most of the cases will fall in this range) one gets inf S1⊕S2 = inf S1. inf S2, sup S1⊕S2 = sup S1⊕sup S2 and, as some particular cases, we have {a} ⊕S2 = {x|x=a⊕s2, where s2∈S2}, with inf {a} ⊕S2 = a * inf S2, sup {a} ⊕S2 = a⊕sup S2.

Division of sets by a number: Let k∈R* then S1/k = {x|x=s1/k, where s1∈S1}.

Neutrosophic logic connectors: One uses the definitions of neutrosophic probability and neutrosophic set operations. Similarly, there are many ways to construct such connectives according to each particular problem to solve; here we present the easiest ones: One notes the neutrosophic logic values of the propositions A1 and A2 by NL (A1) = (T1, I1, F1) and NL(A2) = (T2, I2, F2) respectively.

For all neutrosophic logic values below: if, after calculations, one obtains numbers <0 or >1, one replaces them $^-0$ or 1^+ respectively.

Negation: NL(¬A1) = ({1^+}ϴT1, {1^+}ϴI1, {1+}ϴF1) 1 l

Conjunction: NL (A1^A2) = (T1⊕T2, I1⊕I2, F1⊕F2). (And, in a similar way, generalized for n propositions.)

Implication: NL (A1↔A2) = ({1^+} ϴT1⊕T1⊕T2, {1^+} θI1⊕I1⊕I2, {1^+} ϴF1⊕F1⊕F2).

Neutrosophic relation: A neutrosophic relation R on scheme ∑ is any subset of τ (∑)× [0, 1] × [0, 1]. For any t€τ (∑), we shall denote an element of R as ‹t,R(t)$^+$,R(t)$^-$ ›, where R(t)$^+$ is the belief factor assigned to t by R and R(t)$^-$ is the doubt factor assigned to t by R. Let V(∑) bethe set of all neutrosophic relations on ∑.

Consistent neutrosophic relation: A neutrosophic relation R on scheme ∑ is consistent if R $(t)^{+}$ +R $(t)^{-}$ ≤1, for all t€τ (∑). Let C (∑) be the set of all consistent neutrosophic relations on ∑. R is said to be complete if

$R(t)^{+}$ + $R(t)^{-}$ ≥1, for all t € τ (∑). If R is both consistent and complete, i.e., $R(t)^{+}$ + $R(t)^{-}$ = 1, for all t€τ (∑),then it is a total neutrosophic relation and let T (∑) be the set of all total neutrosophic relations on ∑.

A note on interval mathematics: Dealing with the mathematics of Neutrosphic set theory, the crisp theory of interval mathematics is sometimes useful. In this section, we recollect some basic notions of interval mathematics. For our purpose in this paper, we need to consider intervals of non-negative real numbers only.

Let I_1 = [a,b] and I_2 = [c,d] be two intervals of nonnegative real numbers. A point valued non-negative real number r also can be viewed, for the sake of arithmetic, as an interval [r,r].

Some algebraic operations:

- Interval Addition: $I_1 + I_2 = [a+c, b+d]$
- Interval Subtraction: $I_1 - I_2 = [a-c, b-d]$
- Interval Multiplication: $I_1 * I_2 = [ac, bd]$
- Interval Division: $I_1 \div I_2 = [a/d, b/c]$, when c, d ≠ 0
- Scalar Multiplication : k . I_1 = [ka, kb]

Ranking of intervals: Intervals are not ordered. Owing to this major weakness, there is no universal method of ranking a finite (or infinite) number of intervals. But in real life problems dealing with intervals, we need to have some tactic to rank them in order to arrive at some conclusion. We will now present a method of ranking of intervals, which we shall use in our work here in subsequent sections. We consider a decision maker (or any intelligent agent like a company manager, a factory supervisor, an intelligent robot, an intelligent network, etc.) who makes a pre-choice of a decision parameter β∈ [0,1]. The intervals are to be ranked once the decision-parameter β is fixed. But ranking may differ if
the pre-choice β is renewed.

Definition: _-value of an interval: Let J = [a, b] be an interval. The β-value of the interval J is a non-negative real number J_β, given by J_β = (1- β). a+β.b.

Clearly, 0≤ J_β≤ 1 and for β = 0 J_β= a, which

signifies that the decision-maker is pessimistic and also for β= 1 J_β = b which signifies that the decision-maker is optimistic. For β= 0.5 it is the arithmetic-mean to be chosen usually for a moderate decision.

Comparison of two or more intervals we will do here on the basis of β-values of them. If the value of β is renewed, the comparison results may change. The following definition will make it clear. Now Author is proposing α-Neutrosophic-equality search.

_-Neutrosophic equality search: Consider the

Students database as described in section-1. Consider a normal type of query like Project (Student_Name)Where AGE =approximately 30.

The standard SQL is unable to provide any answer to this query as the search for an exact match for the predicate will fail. The value approximately 30 is not a precise data. Any data of type approximately x, little more than x, slightly less than x, much greater than x etc., are not precise or crisp, but they are Neutrosophic numbers (NN). Denote any one of them, say the neutrosophic number approximately x by the notation I(x). We know that a Neutrosophic number is a Neutrosophic Set of the real numbers. Clearly for every member a ∈ dom (AGE), there is a membership value $t_{I(x)}$ (a) proposing the degree of equality of this crisp number a with the quantity approximately x and a nonmember ship value $f_{I(x)(}$a) proposing the degree of none quality . Thus, in neutrosophic philosophy of samarandech, every element of dom (AGE) satisfies the predicate AGE = approximately 30 up to certain extent and does not satisfy too, up to certain extent. But we will restrict ourselves to those members of dom (AGE) which are α-neutrosophic-equal, the concept of which we will define below. Any imprecise predicate of type AGE = approximately 30, or of type AGE = young (where the attribute value young is not a member of the dom(AGE)), is to be called by Neutrosophic-predicate and a query involving Neutrosophic-predicate is called to be a Neutrosophic-query.

Definition: Consider a choice-parameter α∈[0,1]. A member of a of dom (AGE) is said to be α-Neutrosophic-equal to the quantity approximate x if a∈I_α(x), where I_α(x) is the α-cut of the Neutrosophic number I(x). The degree or amount of this equality is measured by the interval $m_{I(x)}$(a) = [$t_{I(x)}$(a), 1-$f_{I(x)}$(a)]. Denote the collection of all such α neutrosophic-equal members from dom (AGE) by the notation AGE_(x), which is a subset of dom (AGE). If AGE_α (x) is not a null-set or singleton, then the members can be ranked by ranking their corresponding degrees of equality.

Definition: Consider a choice value $\beta \in [0,1]$. At β level of choice, for every element a of AGE_α (x), the truth value $t(p_1, p_2)$ of the matching of the predicate p_1: given by AGE = approximately x with the predicate p_2: AGE = a is equal to the β-value of the interval $m_{I(x)}(a)$.

Neutrosophic- proximity search: The notion of α-neutrosophic-equality search as explained above is appropriate while there is an Neutrosophic-predicate in the query involving NNs. But there could be a variety of vague predicates existing in a Neutrosophic query, many of them may involve Neutrosophic hedges (including concentration/dilation) like good, very good, excellent, too much tall, young, not old, etc. In this section we present another type of search for finding out a suitable match to answer imprecise queries. In this search we will use the theory of neutrosophic-proximity relation[4,5]. We know that a neutrosophic-proximity relation on a universe U is a neutrosophic relation on U which is both neutrosophic-reflexive and neutrosophicsymmetric.

Consider the Students database as described in section-1 and a query like Project (Student_Name) Where Eye-Color = dark-brown.

The value/data dark-brown is not in the set dom (Eye-Color). Therefore a crisp search will fail to answer this. The objective of this research work is to overcome this type of drawbacks of the classical SQL. For this we

notice that there may be one or more members of the set dom (Eye-Color) which may closely match the eye color of brown or dark- brown.

Consider a new universe given by W = dom(EYECOLOR) U {dark-brown}.

Propose a Neutrosophic-proximity relation R over W. Choose a decision-parameter $\alpha \in [0, 1]$. We propose that search is to be made for the match e ∈ dom(EYECOLOR) such that t_R(dark-brown, e)$\geq\alpha$.

(It may be mentioned here that the condition

t_R(dark-brown,e) $\geq\alpha$ does also imply the condition f_R(dark-brown,e) $\leq 1-\alpha$).

We say that e is a close match with dark-brown with the degree or amount of closeness being the interval $m_{\text{dark-brown}}(e)$ given by $m_{\text{dark-brown}}(e)$ = [t_R(dark brown, e), 1- f_R(dark brown,e)].

At β level of choice, the truth-value t (p_1, p_2) of the matching of the predicate p_1: given by EYE-COLOR = dark-brown with the predicate p_2: AGE = e is equal to the β-value of the interval $m_{\text{dark-brown}}(e)$.

Neutrosophic-search: In this section we will now present the most generalized method of search called by Neutrosophic-search. The Neutrosophic-search of matching is actually a combined concept of α- neutrosophic-equality search, neutrosophic-proximity search and crisp search.

For example, consider a query like Project

(Student_Name) Where (Sex = M, Eye-Color = dark-brown, Age= approximately 30).

This is a neutrosophic-query. To answer such a query, matching is to be searched for the three predicates p_1, p_2 and p_3 given by:

- p_1: SEX = M,
- p_2: EYE-COLOR = dark-brown and
- p_3 : AGE = approximately 30

where p_1 is crisp and p_2, p_3 are neutrosophic(imprecise).

Clearly, to answer this query the proposed

neutrosophic search method is to be applied, because in addition to crisp search, both of α-neutrosophic-equality search and neutrosophic-proximity search will be used to answer this query. The truth-value of the matching of the conjunction p of p_1, p_2 and p_3 will be the product of the individual truth values, (where it is needless to mention that for crisp match the truth-value will be exactly 1). There could be a multiple number of answers to this query and the system will display all the results ordered or ranked according to the truth-values of p.

It is obvious that the neutrosophic-search technique for predicate-matching reduces to a new type of fuzzy search technique as a special case.

Conclusion

In this study, we have introduced a new method to answer imprecise queries of the lay users from the databases (details of the databases may not be known to the lay (users). We have adopted Neutrosophic set tool to solve the problem of searching an exact match or a close match (if an exact match is not available) of the predicates so that we will be able to get the answer of evidence for you (i.e., exact/truth match) and evidence against you (i.e., false match) and the undecidability (i.e., indeterminacy) This is a complete new Method of Answering Queries based on Neutrosophic logic.

REFERENCES

1. Gau, W.L. and D.J. Buehrer, 1993. Vague sets. IEEE Trans. Syst., Man and Cybernetics, 23:610-614. ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=5923&isYear=1993&count=32&page=1

2. Atanassov, K., 1986. Intuitionistic fuzzy sets. Fuzzy Sets System, 20: 87-96. linkinghub.elsevier.com/retrieve/pii/016501149400286G

3. Atanassov, K., 2000. Intuitionistic Fuzzy Sets: Theory and Applications. Physica Verlag, New York. ISBN 37908 1425 3

4. Bustince, H. and P. Burillo, 1996. Vague sets are intuitionistic fuzzy sets. Fuzzy Sets Syst., 79: 403-405portal.acm.org/citation.cfm?id=241549

5. Chiang D., L.R. Chow and N. Hsien, 1997. Fuzzy information in extended fuzzy relational databases. Fuzzy Sets Systems, 92: 1-10. www.elsevier.com/locate/fss

6. Barbara, D., H. Garcia-Molina and D. Porter, 1992. The management of probabilistic data. IEEE Trans. Knwl. Data Eng., 4; 487-502.

7. Smarandache, F. 2002. A unifying field in logics: Neutrosophic filed. Multiple Valued Logic Int. J., 8: 385- 438. www.gallup.unm.edu/~smarandache/eBook-Neutrosophics2.pdf

8. Baize, V. and A. Gilio, 2000. A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. Int. J. Approximate Reasoning. Volume 24, Number 2-3, MAY 2000 linkinghub.elsevier.com/retrieve/pii/S0888613X00000384

9. Biazzo V., A. Gilio and G. Sanfilippo, 1999. Efficient Coherence Checking and Propagation of Imprecise Probability Assessments . In Proceedings IPMU-2000.

10. Cavallo, R. and M. Pittarelli, 1987. The theory of probabilistic database. In Proceedings of the 13th VLBDB Conference, Brighton, England, Page numbers 71-78.

11 Codd, E.F., 1979. Extending the database relational model to capture more meaning. ACM Trans. Database Syst., 4: 394-405.

12 Coletti, G., 1994. Coherent numerical and ordinal probabilistic assessments. IEEE Trans. Syst. Man Cybernetics, 24: 1747-1754. www. ieeexplore.ieee.org

13. Coletti, G. and R. Scozzafava, 1996. Characterization of coherent conditional probabilities as a tool for their assessment and extension. J. Uncertainty, Fuzziness Knowledge-Based Syst., 4: 103-127. portal.acm.org/citation.cfm?id=766838

14. G. Coletti and R. Scozzafava, Exploiting zero probabilities, in: Proc. of EUFIT '97, Aachen, Germany (ELITE foundation, 1997) pp. 1499-1503. 5th European Congress on Intelligent Techniques and Soft Computing, September 08. - 11, 1997. www.eufit.org/proceedings/97/volume_2.html

15. Dey, D. and S. Sarkar, 1996. A probabilistic relational model. ACM Trans. Database Syst., 21: 394-405.

16. Re, C., N. Dalvi and D. Suciu, 2007. Efficient topk query evaluation on probabilistic data. In: Proceedings of ICDE(IEEE International Conference on Data Engineering),2007. pages 886–895. www.icde2007.org

Automatic Decipherment of Ancient Indian Epigraphical Scripts - A Brief Review

Soumya A[1] and G Hemantha Kumar[2]

[1]Department of Computer Science & Engineering, R V College of Engineering, Karnataka, India,
[2] Department of Studies in Computer Science, University of Mysore, Karnataka, India,
[1]soumyaa@rvce.edu.in
[2]ghk2007@yahoo.com

Abstract: The history of writing in India dates back to the 3rd millennium BC as is evident from the seals and clay pottery fragments bearing short inscriptions discovered in various parts of India. These seals and various artifacts are known to belong to the ancient civilization of Indus Valley; the mature phase of this civilization is recognized as Harappan Civilization, spanning period between c.3500 and 1700 BC. There are thousands of inscriptions found across various regions in India. Importance of inscriptions to mankind is remarkable. Although the claims of decipherment are made, no acceptable reading of the inscriptions is yet possible. The scripts of modern Indian languages have evolved over centuries. We can observe changes in characters during the phase of evolvement. Many difficulties are faced by modern readers in interpreting an ancient script. To decipher ancient script initially the era to which a given ancient script belong to has to be predicted, followed by automatic recognition of ancient script. This knowledge can be used by archaeologists and historians for further explorations.

Keywords: Inscription, Epigraphy, Paleography, Document Image Analysis, Character Recognition.

1. Introduction

Among many ancient societies, writing held an extremely special and important role. A writing system as a set of visible or tactile signs used to represent units of language in a systematic way. It is true that many non-writing cultures often pass long poems and proses from generation to generation without any change, and writing cultures can't seem to do that. But writing was a very useful invention for complex and high-population cultures. Writing was used for record keeping to correctly count agricultural products, for keeping the calendar to plant crops at the correct time, for religious purpose (divination) and socio-political functions (reinforcing the power of the rulers). In past centuries, scientists had used writing as one of the "markers" of civilization [1].

Scripts denote the writing systems employed by the languages to represent the sounds which form the phonetic base of the language. In India, prior to invention of writing or printing papers, Palmyra leaves and birch leaves were used for writing purposes. As they could not be long lasting, engraving on rocks, pillars and plates made of copper/ gold/ silver came into practice. Epigraphy (derived from two Greek words viz., **epi** meaning **on or upon** and **graphie** meaning **to write)**, is the study of inscriptions engraved on stone or other durable materials, or cast in metal. It is the science of classifying inscriptions according to cultural context and date, elucidating them and assessing what conclusions can be deduced from them. The person studying this is called an **epigrapher or epigraphist**. Many of the inscriptions are couched in extravagant language, but when the information gained from inscriptions can be corroborated with information from other sources such as still existing monuments or ruins, inscriptions provide insight into India's dynastic history that otherwise lacks contemporary historical records [2]. The inscriptions provide valuable information about history, culture, astronomy, medicine, management, political, religious, social, economic, administrative and educational conditions that prevailed during ancient periods.

Many inscriptions do not contain enough historical details to fix their authorship conclusively. For example, from the inscriptions with the name Rajaraja, it is not very clear whether Rajaraja the First or the Second or the Third is intended. To assign dates to such inscriptions and to identify the rulers, palaeography is the main tool. **Paleography** is the study of ancient handwriting and the practice of deciphering and reading historical manuscripts. The paleographer must have the knowledge of: first, the language of the text and second, the historical usages of various styles of handwriting, common writing customs, and scribal/notarial abbreviations.

Language and Script are two different entities. The relation between a language and a script is neither 'original' nor 'fixed'.Any language can be written in any script. Having or not having 'own script' is neither a status nor any hurdle for a language. Three important varieties of scripts that were prevalent in ancient India were: Indus valley script, Brahmi Script and Kharosti script. The scripts of modern Indian languages have evolved from one of these ancient scripts over the centuries [3]. The evolution of the script is dependent on many factors: the writing material, (Stone, Copper, Palm leaf, Paper etc), writing tools, modes of writing and the background of the scribes. Important inventions with advanced technology such as those of paper, printing, typing and the fonts used in computers have had their own influences over a period of time. In India currently there are 13 Scripts and 23 official languages for communication at state level. Apart from these, there are many languages and dialects, used by a number of people.

Scripts have evolved over centuries to the present form In every century each letter was written in a particular style. During the regime of a single ruler, inscriptions may have different features for the same characters. Each inscriber had his individual style and there was a good deal of diversity in style in a given period. There was also a certain amount of regional variation [3], [4], [5]. Thus even for the experts, it is difficult to assign dates to many inscriptions, whether complete or fragmentary It is observed that many of the Indian languages evolved since 3rd century B.C. and the characters have assumed different shapes over the centuries. Modern readers find difficulties in interpreting an ancient script. The expert epigraphists decipher these scripts and translate them into the regional languages. These expert epigraphists are few and it is expected that they could become extinct in near future and also the significance of inscriptions to mankind is enormous. Hence there is a dire need for the automation of deciphering the inscriptions into an understandable form, which would help archaeologists and historians to know the cultural heritage of the civilization, so as to enable further explorations.

The image of inscriptions captured are subjected to various types of degradations like erased characters, broken characters, touching characters, non-uniform spacing of the text between lines and characters, unwanted marks engraved , add complexity in segmenting the text into lines, words or characters. Inturn poor results of segmentation, affect the classification and recognition accuracy. Nevertheless, the classification and recognition of epigraphical document image remains to be one of the most challenging problems in Pattern recognition and Image analysis.

2. Related work

Extensive research has been carried out on Optical Character Recognition (OCR) in the last few decades. Many commercial and accurate systems are now available for machine-printed character recognition. Unfortunately, the success obtained with the machine-printed OCR systems has not readily been transferred to the handwriting recognition arena. High accuracy OCR systems are reported for English with excellent performance in presence of printing variations and document degradation. Recognizing English characters is much simpler as there are only 26 letters and each letter is quite distinct from others compared to recognition of Indian language characters. For Indian and many other oriental languages- OCR systems are not yet able to successfully recognize printed / handwritten document images of varying scripts, quality, size, style and font. Many researchers have been working on script recognition for more than three decades but there are very few tools to identify these scripts. Compared to European languages, Indian languages pose many additional challenges. Indian languages are characterized with the properties: (i) large number of vowels, consonants, and conjuncts. (ii) Have a base character along with vowels attached, forming single character called compound character. (iii) Most scripts spread over several zones. (iv) Lack of standard test databases (ground truth data) of the Indian languages. In India also pioneering work has been done on several scripts like Bangla, Devanagari, Telugu, Tamil, Kannada, etc.

These conventional OCRs address the recognition of characters of various scripts of modern period. It is observed that many of the scripts have been evolved from the Brahmi script which is assumed to be present in 3rd century B.C. Since then, the evolution in scripts has been taken and there are many scripts today. The reported work in the field of developing a computer-based system for recognizing the text of epigraphical documents is very less. There are only few works carried out in this area on Indian script in general. Hence, it is the need of the time for the complete automation of deciphering epigraphical scripts written in olden days. The scripts of modern Indian languages have evolved to the present form over the centuries, leading to changes in characters over a period of time. Hence initially the dating of given input inscription is to be done, so as to have an idea of which character set to be applied for automatic reading of inscriptions Age identification and recognition of ancient epigraphical scripts is a problem under the genre of pattern recognition and image analysis.

Over the last few years, several of the major epigraphic corpora have begun digitization projects. The epigraphic community also hopes to create a unified database of information about all known Greek and Latin inscriptions [6]. A digitized corpus of inscriptions can include several different representations of the inscriptions: photographs of inscriptions; photographs of ‘squeezes’ of inscriptions, which are casts of the stone made in a flexible material like paper or latex; diplomatic transcriptions; edited texts; translations; commentaries. Many projects also find it convenient to store meta-data about the inscriptions in a database, to facilitate searching. The most useful meta-data fields include the date of the inscription, its language, the types of letter forms in use in it, where it was found, what material it is on, and its size.

Lagrange M., and Renaud, H have simulated reasoning by means of an expert system in archaeology from France using computer [7].

A few projects linking Indian epigraphy with the computer technology have been proposed and implemented with a fairly high degree of success. Siromoney, G [1975] has demonstrated the use of computer techniques for enhancement of an image and information retrieval for reading ancient Tamil inscriptions [8]. A statistical analysis of personal names in ancient Indian inscriptions has been reported by Karashima and Subbarayalu [1976]. Siromoney G., Chandrasekaran M. and Chandrasekaran R [1981] have shown that, to assign approximate date to ancient and medieval Tamil inscriptions of unknown authorship found in the southern part of India and recognition of an ancient Tamil script of the Chola period Indian Script, computer techniques may be used [8]. Chandrasekaran, R. [1982] has worked on recognition of certain ancient and modern Indian scripts using computer techniques [8]. A work on automated recognition of ancient Indian Scripts in general and ancient Brahmi script in particular by Anasuyadevi [2000] has been reported. She has proposed a fuzzy neural network for the recognition of Brahmi characters [17].

Concept of Component analysis is applied to the study of South Indian sculpture. An expert system was developed for Indian epigraphy used to assign probable dates to medieval Tamil inscriptions [18]. The program was developed in BASIC on a Genie-1 Microcomputer [1985].

K Harish Kashyap, Bansilal, P Arun Koushik [2000] have proposed a hybrid neural network architecture for age identification of ancient Kannada scripts, which focuses on classification and age identification of different characters by a hybrid model. After pre-processing the characters, the work is implemented in two phases. The first phase- identifies the base character, incorporates an Artificial Neural Network (ANN). ANN is trained by Back propagation algorithm to identify the present day base character corresponding to input character. In the second phase - for identification of age pertaining to the base character, a Probabilistic Neural Network (PNN) - a Bayesian classifier is used, taking the advantage that no training is involved prior to classification [19].

The research work carried out by Srikanta Murthy K [2005] provides novel methods for preprocessing - for removal of noises, segmentation of lines and characters, thinning and finally classification of the epigraphical documents belonging to different periods. The work aims at transformation of the epigraphical object into an image of readable form. Two preprocessing techniques for removal of noise have been proposed – the first algorithm is based on a rectangle fitting wherein the height of the character to be retained is assumed to be greater than the noisy pixel. The second algorithm employs a template to obtain the minimum majority of noisy pixels. Segmentation of lines and characters are carried out using- a Partial Eight Direction Based Line Segmentation (PEBLS) algorithm wherein horizontal Projection profile is applied to identify the base and supplementary reference lines. The second approach is based on Nearest Neighbor Clustering (NNC), which could be used even when the document is skewed. Three thinning algorithms - two-step algorithm, fully parallel thinning algorithm and rotation invariant four- step algorithm have been designed. Classification of the epigraphical document belonging to different period is carried out using - a method based on texture features, a method based on invariant moments which are invariant to rotation, translation and scaling, and for accurate estimation of the period, a neural network based approach is adopted [20].

3. Overview of the system for deciphering ancient scripts

The complete automation of classification and recognition of ancient epigraphical scripts involves the following steps and the workflow is as shown in Figure 1.

- The input image of the inscription may be degraded due to the presence of the broken characters, erased characters, touching characters, distortion due to fossils settled, irrelevant symbols engraved by the scribes and so on. Also the non uniform spacing between the lines and characters of epigraphical document and the skew could complicate the process of deciphering the script. Input epigraphical image has to be subjected to pre-processing stage initially. Hence suitable preprocessing techniques for removal of noise and segmentation of lines and characters are to be devised.

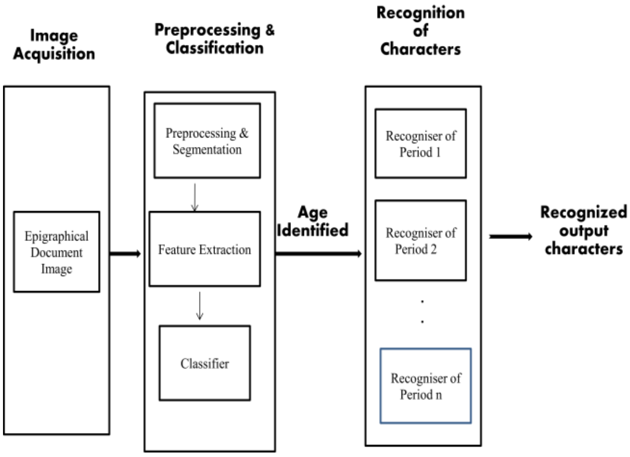


Image Acquisition
Preprocessing & Classification
Recognition of Characters
Epigraphical Document Image
Preprocessing & Segmentation
Feature Extraction
Classifier
Age Identified
Recogniser of Period 1
Recogniser of Period 2
Recogniser of Period n
Recognized output characters

Figure1. System architecture for Classification and Recognition of ancient epigraphical scripts

- Features have to be extracted for the segmented characters, so that the task of classifying the pattern is made easy by a formal procedure. Appropriate feature extraction methods have be devised for measuring the relevant shape information contained in a pattern.
- Characters have evolved over centuries to the current form undergoing several twins and turns. Different periods have different character set. Hence the period of epigraphical script has to be predicted so as to know which character set has to be used for supervisory reading of ancient epigraphy documents.
- Finally for automatic decipherment of ancient epigraphical scripts, recognizers are to be devised which takes the epigraphical document image of ancient script, whose period has been predicted as the input and outputs the text in a readable form.

Hence we need to seek new approaches for transforming ancient epigraphical script into recognizable form.

4. Conclusion

The epigraphical survey is of importance as inscriptions provide insight into history of the region during various dynasties which otherwise lacks historical records. It helps many scholars who are working in the field of history, archaeology and linguistics. Since the contribution of the inscriptions to the society is remarkable and the expert epigraphists could become extinct in future, a complete automated system with sufficient intelligence has to be developed to decipher the epigraphical documents. To sum up, the research issue addressed here is to produce a computer perceivable image from a raw epigraphical script which are the inscriptions on rocks or pillars or plate, then classification of the ancient script into respective periods and recognition of the characters which would assist historians and archeologists to know the cultural heritage of the civilization so as to enable further exploration.

References:

[1] Writing Systems: tttp://www.ancientscripts.com/ws.html

[2] Possehl Gregory L. Indus Age: The writing System, University Pennsylvania Press, Philadelphia, (1996).

[3] A.V.Narasimha Murthy, "Kannada Lipiya Ugama Mattu Vikasa", Kannada Adhyayana Samsthe, Mysore University, Mysore, (1968).

[4] Dr M G Manjunath, G K Devarajaswamy, "Kannada Lipi Vikasa, Jagadhguru", Sri Madhvacharya Trust, Sri RagavendraSwamy Matt, Mantralaya.

[5] Dr. Devarakonda Reddy, "Lipiya Huttu Mattu Belavanige — Origin and Evolution of Script", Published by Kannada Pustaka Pradhikara (Kannada Book authority), Bangalore.

[6] Electronic Textual Editing: Epigraphy [Anne Mahoney, Perseus Project & Stoa Consortium Tufts University] http://www.tei-c.org/About/Archive_new/ETE/Preview/mahoney.xml#body.1_div.3

[7] Lagrange, M., and Renaud, H, "Intelligent knowledge-based systems in archaeology: a computerized simulation of reasoning by means of an expert system", Computers and the Humanities, Vol.19, pp. 37-49, (1985).

[8] Works on Epigraphy [online]: http://Dr Gift Siromoney/epigraphy.

[9] Siromoney, G., "Computer techniques of image enhancement in the study of Pallava Grantha inscription", Studies in Indian Epigraphy 2, pp. 55-58, (1975).

[10] Siromoney, G., Chandrasekaran, M. and Chandrasekaran, R., "Computer methods of dating medieval Tamil inscriptions", STAT-26/76, the Third Annual Congress of the Epigraphical Society of India at Udupi (March 1978).

[11] Siromoney, G., Chandrasekaran, M. and Chandrasekaran R, "Computer recognition of an ancient Tamil script of the Chola period", Journal of the Epigraphical Society of India, Vol. VI, pp 18-19, (1978).

[12] Siromoney, G., Chandrasekaran, M. and Chandrasekaran R, "Computer recognition of an ancient common Indian Script", STAT-36/78, the Symposium on the Use of Indian Languages in Computer based Information Systems, (March 1978).

[13] Siromoney, G., M. Bagavandas and S. Govindaraju, "An application of component analysis to the study of South Indian sculpture", Computers and the Humanities 14, pp. 29-37, (1980).

[14] Siromoney, G., M. Chandrasekaran and R. Chandrasekaran, , "Computer methods of dating Tamil inscriptions" , Proceedings of the Fifth International Conference-Seminar of Tamil Studies, Madurai, India, pp. 2.7-2.13, (1981).

[15] Chandrasekaran, R., "Computer recognition of certain ancient and modern Indian script"', Ph.D. Thesis, University of Madras, (1982).

[16] Siromoney, G., M. Chandrasekaran and R. Chandrasekaran, "Computer dating of medieval inscriptions: South Indian Tamil", Computer and the Humanities, Vol. 17, pp. 199-208, (1983).

[17] Anasuya Devi H.K, , "Automated Recognition of Ancient Indian Scripts", Proceedings of National workshop on Computer Vision, Graphics and Image processing, WVGIP, pp 216-219, (2002).

[18] Gift Siromoney, Chandrasekaran R. and Suresh D., "Developing an expert system for Indian epigraphy" at the Kibble Center for Statistical Computing at the Department of Statistics, Madras Christian College, (1985).

[19] K Harish Kashyap, Bansilal, P Arun Koushik, "Hybrid Neural Network Architecture for Age Identification of Ancient Kannnada Scripts", Proceedings of 2003 International Symposium on Circuits and Technology (ISCAS 2003), Vol 5, Pg V661- V664, (2003).

[20] Srikanta Murthy.K, "Transformation Of Epigraphical Objects Into Machine Recognizable Image Patterns", Ph.D Thesis, University Of Mysore, Mysore, (December 2005).

[21] Works on OCR for Indian Languages [online] http:// Indira Gandhi National Centre for the Arts (IGNCA's) Southern Regional Centre, Bangalore.

[22] S Pletschacher, J Hu and A Antonacopoulos, "A New framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering",10th

International Conference on Document Analysis and Recognition, (2009).

[23] Sheikh Faisal Rashid, Faisal Shafait and Thomas M. Breuel, "Connected Component level Multiscript Identification from Ancient Document Images", (2010).

[24] D Dayalan, "Computer Application in Indian Epigraphy", Bharatiya Kala Prakashan publication, (2005).

Author Biographies

Soumya A, Assistant Professor, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore. She has obtained M.S degree in Computer Cognition Technology, University of Mysore, Mysore and B.E in Computer Science and Engineering, Bangalore University, Bangalore. Her areas of research are Artificial Intelligence, Soft Computing, Pattern Recognition and Image Processing. She has guided several under graduate projects and 7 post graduateprojects.

Dr G Hemantha Kumar, Professor & Chairman, Department of studies in Computer Science, University of Mysore, Mysore. Serving as Course -Coordinator of Chinese B.Tech Programme, University of Mysore, Mysore. He was awarded Ph.D. for his thesis titled: "On Automation of Text Production from Pitman Shorthand Notes" from University of Mysore, Mysore. His areas of research are: Image Processing, Pattern Recognition, Numerical Techniques, and Bio-Metric. He has to his credits 31 publications in International / National Journals and 46 publications in International / National Conferences/Workshops. He has guided several PhD candidates and is presently guiding 5 PhD candidates.

An Analysis of Software Cost Estimation Traditional Models with Neural Network Based Approach

Manpreet Kaur*, Sushil Garg*, Mamta Sharma

*Dept CSE, RIMT-IET, Mandigobindgarh, Fatehgah Sahib

manumtech@yahoo.com, mmta9976@gmail.com, sushilgarg70@yahoo.com

Abstract: *Software effort estimation actually encompasses all estimation, risk analysis, scheduling, and SQA/SCM planning. However, in the context of set of resources, planning involves estimation - your attempt to determine how much money, how much effort, how many resources, and how much time it will take to build a specific software-based system or product. In this paper we will study the efficiency of Neural Network based cost estimation model with the traditional cost estimation model like Halstead Model, Bailey-Basili Model, Doty Model. We conclude our result with the proposal of Neuron based Model basis on Back propagation Technique.*

1. Introduction

A Neural Network (NN) is a computer software (and possibly hardware) that simulates a simple model of neural cells in animals and humans. The purpose of this simulation is to acquire the intelligent features of these cells. In this document, when terms like neuron, neural network, learning, or experience are mentioned, it should be understood that we are using them only in the context of a NN as computer system. NNs have the ability to learn by example, e.g. a NN can be trained to recognize the image of car by showing it many examples of a car.

2. Literature Survey

Accurate estimate means better planning and efficient use of project resources such as cost, duration and effort requirements for software projects especially space and military projects [1], [2]. Efficient software project estimation is one of the most demanding tasks in software development. Problem of inaccurate estimate for projects and in many cases inability to set the correct release day for their software correctly lead to inefficient use of project resources. Unfortunately, software industry suffers the problem of incorrect estimate for projects and in many cases inability to set the correct release day for their software correctly. This leads to many losses in their market, e.g. risk due to low quality of the deliverables and penalties for missing the deadlines. Normally, estimation is performed using only human expertise [3], [4], but recently attention has turned to a variety of computer-based learning techniques.

In 1995, Standish Group served over 8,000 software projects for the purpose of budget analysis. It was found that 90% of these projects exceeded its initially computed budget. Moreover, 50% of the completed projects lake the original requirements [5]. From these statistics, it can be seen how prevalent the estimation problem is. Evaluation of many software models were presented in [6], [7], [8].

Numerous models were explored to provide better effort estimation [9], [10], [11], [12]. In [4], [13], authors provided a survey on the effort and cost estimation models.

Serious research in the Neural Network area is started in the 1950’s and 1960’s by researchers like Rosenblatt (Perceptron), Widrow and Hoff (ADALINE). In 1969 Minsky and Papert wrote a book exposing Perceptron limitations. This effectively ended the interest in neural network research. In the late 1980’s interest in NN increased with algorithms like Back Propagation, Cognitrons and Kohonen. (Many of them where developed quietly during the 1970s)

In the literature of Neural Networks (NNs) The following function is called a Sigmoid function.:

$$s(x)= 1/ (1 + e^{-a * x})$$

The coefficient a is a real number constant. Usually in NN applications a is chosen between 0.5 and 2. As a starting point, you could use a=1 and modify it later when you are fine-tuning the network. Note that s(0)= 0.5, s(∞)= 1, s(-∞)=0. (The symbol ∞ means infinity).

The Sigmoid function is used on the output of neurons. In a NN context, a neuron is a model of a neural cell in animals and humans. This model is simplistic, but as it turned out, is very practical. In NN the inputs simulate the stimuli/signals that a neuron gets, while the output simulates the response/signal which the neuron generates. The output is calculated by multiplying each input by a different number (called weight), adding them all together, then scaling the total to a number between 0 and 1.

The following diagram shows a simple neuron with:

1. Three inputs $[x_1, x_2, x_3]$. The input values are usually scaled to values between 0 and 1.
2. Three input weights $[w_1, w_2, w_3]$. The weights are real numbers that usually are initialized to some random numbers. Do not let the term weight mislead you, it has nothing to do with the physical sense of weight, in a programmer context, think of

the weight as a variable of type float/real that you can initialize to a random number between 0 and 1.

3. One output is shown as z. A neuron has one (and only one) output. Its value is between 0 and 1. It can be scaled to the full range of actual values.

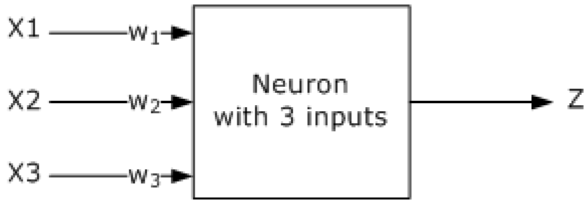


Fig. 1 Neuron Model with 3 inputs

Let

$$d= (x_1 * w_1) + (x_2 * w_2) + (x_3 * w_3)$$

In a more general fashion, for n number of inputs:

$$d = \sum_{i=1}^{n} x_i w_i$$

Let θ be a real number which we will call Threshold. Experiments have shown that best values for θ are between 0.25 and 1. Again, in a programmer context, θ is just a variable of type float/real that is initialized to any number between 0.25 and 1. When sigmoid function, s( ), is applied:

$$z= s(d + \theta)$$

This says that the output z is the result of applying the sigmoid function on (d + q). In NN applications, the challenge is to find the right values for the weights and the threshold.

The following diagram shows a Back Propagation NN:

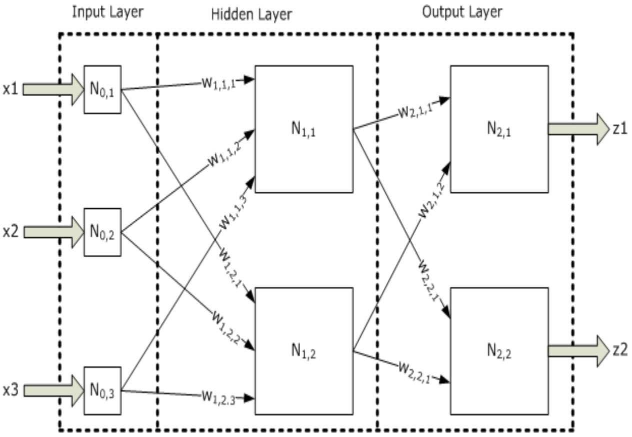


Figure 2: Back Propagation Network

The above NN consists of three layers:

- Input layer with three neurons.
- Hidden layer with two neurons.
- Output layer with two neurons.

The output of a neuron in a layer goes to all neurons in the following layer. Each neuron has its own input weights. The weights for the input layer are assumed to be 1 for each input. In other words, input values are not changed and the output of the NN is reached by applying input values to the input layer, passing the output of each neuron to the following layer as input.

The Back Propagation NN must have at least an input layer and an output layer. It could have zero or more hidden layers.

The number of neurons in the input layer depends on the number of possible inputs we have, while the number of neurons in the output layer depends on the number of desired outputs. The number of hidden layers and how many neurons in each hidden layer cannot be well defined in advance, and could change per network configuration and type of data. In general the addition of a hidden layer could allow the network to learn more complex patterns, but at the same time decreases its performance. You could start a network configuration using a single hidden layer, and add more hidden layers if you notice that the network is not learning as well as you like e.g. suppose we have a bank credit application with ten questions, which based on their answers, will determine the credit amount and the interest rate. To use a Back Propagation NN, the network will have ten neurons in the input layer and two neurons in the output layer.

The Back Propagation NN works in two modes, a supervised training mode and a production mode. The training can be summarized as follows:

First, start by initializing the input weights for all neurons to some random numbers between 0 and 1, then:

i. Apply input to the network.
ii. Calculate the output.
iii. Compare the resulting output with the desired output for the given input. This is called the error.
iv. Modify the weights and threshold q for all neurons using the error.
v. Repeat the process until error reaches an acceptable value (e.g. error < 1%), which means that the NN was trained successfully, or if we reach a maximum count of iterations, which means that the NN training was not successful.

A suitable training algorithm can be used for updating the weights and thresholds in each iteration (step IV) to minimize the error.

Changing weights and threshold for neurons in the output layer is different from hidden layers. Note that for the input layer, weights remain constant at 1 for each input neuron weight.

The literature considered the mean magnitude of relative error (MMRE) as the main performance measure.

The value of an effort predictor can be reported many ways including MMRE. MMRE value is computed from the relative error, or RE, which is the relative size of the difference between the actual and estimated value:

RE.i = (estimate.i - actual.i) / (actual.i)

Given a data set of of size "D", a "Training set of size "(X=|Train|) <= D", and a "test" set of size "T=D-|Train|", then the mean magnitude of the relative error, or MMRE, is the percentage of the absolute values of the relative errors, averaged over the "T" items in the "Test" set; i.e.

MMRE.i = abs(RE.i)

MMRE = 100/T*(MRE.1 + MRE.2 + ... + MRE.T)

The mean magnitude of relative error (MMRE) can also be written as:

$$MMRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}$$

Where y_i represents the i^{th} value of the effort and $\hat{y}_i$ is the estimated effort.

The another evaluation criteria to measure the performance of the developed models using n measurements selected to be the route mean of the sum square of the error:

$$RMSSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y_i represents the ith value of the effort and $\hat{y}_i$ is the estimated effort.

3. Result & Discussion

The dataset of [10] is used for the comparison of different models. In this dataset, there is empirical data in terms of KLOC, Function Point and Effort values of 18 projects as shown in table I.

The data of first 13 projects is used as training data for the Neural Network and data of last 5 projects is used as testing data of the trained Neural Network. The neural network used is backpropagation based Neural Network that consists of two neurons in input layer, two neurons in the hidden layer and one neuron in the output layer. In the testing phase the calculated efforts and errors using different models is shown in table 1 and table 2 respectively.

Table1. Data of Actual Effort Required

Project No.	KLOC	Function Point	Actual Effort (in person-hour)
1	95.2	31	125.8
2	50.2	21	90
3	56.5	22	81
4	56.5	21	91.8
5	32.1	38	42.6
6	67.5	29	98.4
7	15.8	29	20.9
8	10.5	34	10.3
9	21.5	31	28.5
10	5.1	29	9
11	4.2	17	8
12	9.8	32	8.3
13	22.1	38	6
14	7	29	8.9
15	88.6	45	100.7
16	10.7	32	17.6
17	13.5	29	25.9
18	105.8	39	148.3

4. conclusion

The performance of the Neural Network based effort estimation system and the other existing Halstead Model, Walston-Felix Model, Bailey-Basili Model and Doty Model models is compared for effort dataset available in literature [15]. The results show that the Neural Network system has the lowest MMRE and RMSSE values i.e. 12.657 and 18.587 respectively. The second best performance is shown by Bailey-Basili software estimation system with 21.385 and 25.1345 as MMRE and RMSSE values. Hence, the proposed Neuro based system is able to provide good estimation capabilities. It is suggested to use of Neuro based technique to build suitable generalized type of model that can be used for the software effort estimation of all types of the projects.

Table 2: Error Calculated In Various Efforts Estimation Models

Perform-ance Criteria	Model Used			
	NN System	Halstead Model	Bailey-Basili Model	Doty Model
MMRE	12.657	155.645	21.385	302.5023
RMSSE	18.587	318.718	25.1345	299.4742

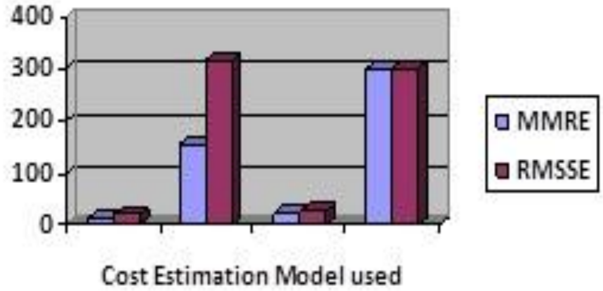


Fig 3 Comparative Analysis of different Cost Estimation Models

References:

1. L. C. Briand, K. E. Emam, and I. Wieczorek, “Explaining the cost of european space and military projects,” in ICSE ’99: Proceedings of the 21st international conference on Software engineering, (Los Alamitos, CA, USA), pp. 303–312, IEEE Computer Society Press, 1999.
2. “Estimating software projects,” SIGSOFT Softw. Eng. Notes, vol. 26, no. 4, pp. 60–67, 2001.
3. J. W. Park R, W. Goethert, “Software cost and schedule estimating: A process improvement initiative,” tech. report, 1994.
4. M. Shepper and C. Schofield, “Estimating software project effort using analogies,” IEEE Tran. Software Engineering, vol. 23, pp. 736–743, 1997.
5. T. S. Group, CHAOS Chronicles. PhD thesis, Standish Group Internet Report, 1995.
6. M. Boraso, C. Montangero, and H. Sedehi, “Software cost estimation: An experimental study of model performances,” tech. report, 1996.
7. O. Benediktsson, D. Dalcher, K. Reed, and M. Woodman, “COCOMO based effort estimation for iterative and incremental software development,” Software Quality Journal, vol. 11, pp. 265–281, 2003.
8. T. Menzies, D. Port, Z. Chen, J. Hihn, and S. Stukes, “Validation methods for calibrating software effort models,” in ICSE ’05: Proceedings of the 27th international conference on Software engineering, (New York, NY, USA), pp. 587–595, ACM Press, 2005.
9. S. Chulani, B. Boehm, and B. Steece, “Calibrating software cost models using bayesian analysis,” IEEE Trans. Software Engr., July-August 1999, pp. 573–583, 1999.
10. B. Clark, S. Devnani-Chulani, and B. Boehm, “Calibrating the cocomo ii post-architecture model,” in ICSE ’98: Proceedings of the 20th international conference on Software engineering, (Washington, DC, USA), pp. 477–480, IEEE Computer Society, 1998.
11. S. Chulani and B. Boehm, “Modeling software defect introduction and removal: Coqualmo (constructive quality model),” tech. report.
12. S. Devnani-Chulani, “Modeling software defect introduction,” tech. report.
13. G. Witting and G. Finnie, “Estimating software developemnt effort with connectionist models,” in Proceedings of the Information and Software Technology Conference, pp. 469–476, 1997.
14. K. Peters, “Software project estimation,” Methods and Tools, vol. 8, no. 2, 2000.

Content Based Recommender Systems

[1] P. Deivendran M.Tech, [2], Dr. T. Mala Ph.D, [3] B.Shanmugasundaram, M.C.A

[1] Assistant Professor, Department of Computer Applications.
Velammal Engineering College, Chennai.600 066, Tamilnadu, India
E-mail: deivendran77p@yahoo.com

[2] Assistant Professor, Department of Information and communication Engineering
Anna University, Chennai – 600 025, Tamilnadu, India
E-mail: malal@cs.annauniv.edu

[3] Assistant Professor, Department of Computer Applications.
Veltech Multitech Dr.Rangarajan Dr.Sakunthula Engineering College,
Avadi, Chennai.600 062, Tamilnadu, India
E-mail: ten1107@gmail.com

Abstract : *Recommender systems help users to identify particular items that best match their interests or preferences. In this paper, we introduce our approach to recommendation based on Case-Based Reasoning (CBR). CBR is a paradigm for learning and reasoning through experience, based on human reasoning. We present a user model based on cases in which we try to capture both explicit interests (the user is asked for information) and implicit interests (captured from user interaction) of a user on a given item. When we apply CBR to recommender systems, some problems arise such as the adaptation of user profiles according to their interests and preferences over time or the utility problem. In order to cope with these problems, our approach includes a "forgetting mechanism" based on the drift attribute. Other systems have implemented CBR approaches to commendation, but unfortunately, only a few evaluate and discuss their results scientifically. This paper also proposes an evaluation technique based on a combination of real user profiles and a user simulator. The results of the simulations show that the forgetting mechanism produces an increase in precision, a decrease in recall and an important reduction of the number of cases in case bases.*

Keywords *– Metrics, Measure, Case Based Reasoning, Cycle, Profile, and Precision.*

1. Introduction

In the real world, making a selection from the incredible number of possibilities the market offers us indeed a laborious work. The main function of the assistants is to advise you. In order to do this, first of all, they have to learn your tastes, interests and preferences. Then, their task consists of looking for information and analyzing[5] the market in order to find out things that may interest you. Since personal assistants are always in contact with you, they also notice your changing interests over time. If you cease to be interested in a certain thing, your personal assistant takes note and finds out what you are presently interested in. Recommender systems draw on previous results from machine learning and other AI technology advances. Among the various machine-learning technologies, we concentrate on Case-Based Reasoning (CBR) as a paradigm for learning and reasoning through experience, as personal assistants do. The main idea of CBR is to solve new problems by adapting the solutions given for old ones. However, when we apply CBR to recommender systems, there are two things missing.

Humans have a vast store of experience on which to base their decisions. When a new problem comes up, humans look for similar problems and try to solve it based on the most similar experiences. However, the time dimension is also present in the human reasoning process. It means that humans have in mind the most recent cases and give them the greater importance when making a decision. When we are dealing with human interests and preferences, the relevance of the most recent cases becomes even more important.

2. Case-Based Recommendation Framework

The core of CBR is a case base which includes all the previous experiences that can give us information we can use to deal with new problems. Then, through the similarity concept, the most similar [12] experiences are retrieved. However, similarity is not a simple or uniform concept. Similarity is a subjective term that depends on what one's goals are. For instance, two products with the same price would get maximum similarity if the user was interested in products with that same price, but would get very different similarity for other concepts, such as quality [11] or trademark. In our approach, the case base represents the user profile and consists of a set of previous experiences (cases); that is, items explicitly and/or implicitly assessed by the user. Each case contains the item description (attributes describing a restaurant in the example) and the interest attributes describing the interests of the user concerning the item. These latter attributes can be explicitly given by the user or implicitly captured by the system.

This kind of recommendation based on similar items is our approach to content-based filtering. With regard to the CBR cycle, we reassess the different phases as follows:

a. In the retrieval phase, i.e. a new item, the system searches for similar items in the case base in order to find out whether the user might be interested in them. Local similarity measures are based on item attributes [14].

b. In the reuse phase, i.e. the retrieved set of similar items, the system calculates a confidence value of interest to recommend the new item to the user based on explicit and implicit interests and the validity of the case according to the user's current interests [8].

c. In the revision phase, i.e. the relevance feedback of the user, the system evaluates the user's interest in the new item. The idea is to track user interaction with the system to get to know relevant information about the user's interest in the recommended item, as well as explicit and implicit information, in order to retain the new case.

d. In the retain phase, the new item is inserted in the case base with the interest attributes that were added in the revision phase. In order to control the case base size, it is also important to know if the user ever gives new feedback [6] about items in the case base. In such a case, it is necessary to forget these interests with time. We propose the use of a new attribute that we call the drift attribute, which will be aware of such changes in user preferences and contribute to case maintenance. In the following sections the structure of the case base and the different CBR phases of the new approach are detailed.

3. Evaluation Metrics

A set of metrics [6] are proposed in order to evaluate recommender systems: precision, recall, measure, fallout, cases, diversity and accuracy.

3.1 Precision

The Precision measure is the fraction of the selected items which are relevant to the user's information need. It is also a measure of selection effectiveness and represents the probability [4] that a selected item is relevant. Precision is calculated with the following formula:

$$P = \frac{s}{n}$$

Where *s* is the number of successful recommendations and *n* is the number of recommendations. The result is a real value ranging from 0 to 1. Precision [9] can also be seen as the probability that a recommendation be successful.

3.2 Recall

The Recall measure is the fraction of the actual set of relevant items which have been correctly classified as relevant. It is a measure of selection effectiveness and represents the probability [13] that a relevant document will be selected. It is interesting to evaluate the number of recommendations that the system makes, since; of course, a recommendation algorithm that recommends all the items will obtain all the possible successes. Recall is computed as follows:

$$R = \frac{n}{t}$$

Where *n* is the number of recommendations and *t* is the total number of possible recommendations. The result of this formula is a real number ranging from 0 to 1. Recall can also be seen as the probability that an item be recommended.

3.3 F-Measure

It is, on occasion, important to evaluate precision and recall in conjunction, because it is easy to optimize either one separately [15]. The F-Measure consists of a weighted combination of precision and recall which produces scores ranging from 0 to 1. When recall increases, precision decreases. Weighting measure between precision and recall called the f-measure. However, we have used a variation of this measure, where the weights[9] are controlled by a parameter *b* [4]. This new approach is calculated as follows:

$$FM = \frac{(b^2 + 1) * P * R}{b^2 * P + R}$$

Where *P* is precision, *R* is recall and *b* is the weighting factor [1]. For example, b = 0.0 means that *FM = precision*; *b = unlimit* means that *FM = recall*; *b* = 1.0 means that recall and precision are equally weighted; *b* = 0.5 means that recall is half as important as precision; and *b* = 2 means that recall is twice as important as precision. We can also see this measure as a modification of precision by recall.

3.4 Fallout

The Fallout measure is the fraction of the non-relevant items selected. It is a measure of rejection effectiveness. We use Fallout to evaluate the percentage of failed recommendations. It is computed like precision, but instead of measuring the recommendations successfully evaluated by the user, we take into account the number of recommendations that the user has valuated as bad. Fallout is calculated with the following formula:

$$F = \frac{u}{n}$$

Where *u* is the number of failed recommendations and *n* is the number of recommendations. Fallout can also be seen as the probability[5] that a recommendation be a failure. The result is a real value confined to the [0-1] interval, although fallout charts represent F normalized between 0 and 100. A fallout value close to 0 means that the system never recommends bad choices; a fallout value of 1 means that the system is always recommending uninteresting items to the user.

3.5. N Cases

The study of the average number of items (cases) contained in the user profile (case base) over time is very important, since it is desirable to reduce the size of the user profiles (solving the utility problem) while preserving or even increasing precision (while adapting the profile to the user). Certainly, the forgetting mechanism will reduce the time and the capacity needed by the algorithms to perform a recommendation.

Thus, Cases is calculated as follows:

$$NC = \frac{\sum_{i=0}^{k} |NC_i|}{k}$$

Where *NCi* is the number of items at the moment *i*, and *k* is the number of moments. That is, the simulation time has been split into *k* units and, in each unit, the number of cases in the case base *NCi* has been measured. At the end of the simulation, the average is computed. Cases is not normalized, therefore, this number is relative to the total number of possible recommendations. What we want to study is the difference between the different Cases from the point of view of different parameters that the forgetting mechanism depends on.

3.6 Diversity

How the reduction of the number of items contained in the user profile affects the diversity within the resulting profiles is an interesting phenomenon for study. To evaluate the diversity, we propose using a well-known clustering method that calculates the number of groups of similar items contained in the profile. The clustering algorithm that we have implemented belongs to a particular subset of clustering methods knows as SAHN[11]: Sequential, Agglomerative, Hierarchical and Non-overlapping methods. The proposed algorithm can be summarized as follows:

3.7 Evaluation Methods

STEP 0: Construction of an initial similarity matrix that contains the pair wise Measures of proximity between the different items of the user profile.

STEP 1: Selection of the two items that are most similar. These alternatives will form a new cluster.

STEP 2: Modification of the similarity matrix creating a cluster with the selected items and recalculating the similarity between the new cluster and the remaining objects. Similarity is calculated with an Arithmetic Average criterion where the similarity between a given item and the cluster is the average similarity between the items composing the cluster and the given item.

STEP 3: Repeat steps 1-2 until the two most similar items have a similarity value over a threshold α. This threshold has to be defined previously, taking into account that it determines the abstraction level achieved. Increasing the threshold we obtain a smaller number of wider (more general) clusters [13]. The number of clusters obtained after the execution of the proposed algorithm is the diversity measure that allows system simulations performed with different parameters to be compared. Thus, a key task is to select a suitable α. Depending on this parameter, the number of clusters constituting the user profile will change. A low α means that only the most similar cases join up and, therefore, the algorithm gives a high number of clusters.

4. Profile Discovering

In order to solve all the shortcomings of the current techniques while benefiting from their advantages, we propose a method of results acquisition called "the profile discovering procedure". This technique can be seen as a hybrid approach between real or laboratory evaluation, log analysis and user simulation. First of all, it is necessary to obtain as many real user profiles as possible. These profiles must contain subjective assessments of the items (preferably explicit evaluations of the user, although the implicit information obtained from the user interaction with the system is also useful). It is desirable to obtain these user profiles through a real or laboratory evaluation although it implies a relatively long period of time. However, it is also possible and faster to get the user profiles through a questionnaire containing all the items which the users have to evaluate. Once the real user profiles are available, the simulation process; that is the profile discovering procedure starts. It consists on the following steps:

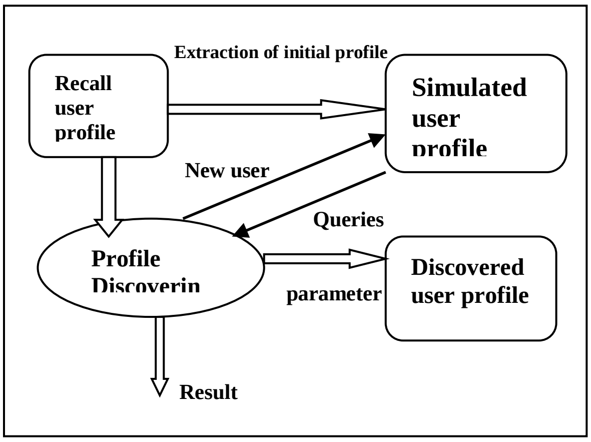
a. Generation of an initial user profile (*UP*) from the real user profile (*RUP*, $UP \subset RUP$).

b. Emulation of the real recommendation process, where a new item (*r*) is recommended From the *UP[9]*.

c. Validation of the recommendation: Otherwise, *r* is rejected.

d. Repeat 2 and 3 until the end of the simulation.

As in the real evaluation, the simulation process starts with the generation of an initial user profile[2]. It is desirable to initially know as much as possible from the user in order to provide satisfactory recommendations from the very beginning. Analyzing the different initial profile generation techniques, namely: manual generation, empty approach, stereotyping and training set, we found different advantages and drawbacks. In manual generation, the user tailors his or her own profile.



(Fig. 1 User Profile Structure)

5. Conclusion

This paper has focused on the study of recommender systems. In particular, we have proposed a recommender system consisting of collaborative recommender agents based on case-based reasoning (CBR) and trust. The CBR cycle has been redefined in order to perform the recommendation task. Assuming that the user has similar interests in similar items, the recommender system predicts the 195 user preferences in new items from the implicit/explicit interest given by the user in similar items.

6. Future Work

The design of a recommender system involves the consideration of a wide range of questions. In addition to the different solutions which have been adopted and described in this paper, many ideas have been proposed, discussed and finally rejected. On the other hand, other questions have remained as undeveloped ideas, which need to be analyzed further and worked on in depth in future work.

References

[1] I. F. A. Iamnitchi, M. Ripeanu. Small-world file-sharing communities. In The 23rd Conference of the IEEE Communications Society (InfoCom 2004), Hong Kong, 2004.

[2] G. Ruffo, R. Schifanella, E.Ghiringhello A Decentralized Recommendation System based on Self-Organizing Partnerships In IFIP-Networking’06, May 2006, LNCS 3976:618-629, Coimbra (Portugal), 2006.

[3] R. L. Jun Wang, Johan Pouwelse and M. R. J. Reinders. Distributed collaborative filtering for peer-to-peer file sharing systems. In Proc. of the 21st Annual ACM SAC, New York, NY, USA, 2006. ACM Press.

[4] B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. AI Magazine, 18(2):37–45, 1997.

[5] K. Lang. NewsWeeder: learning to filter netnews. In Proc. of the 12th ICML, pages 331–339. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995.

[6] M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45:167, 2003.

[7] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill webert: Identifying interesting web sites. In AAAI/IAAI, Vol. 1, pages 54–61, 1996.

[8] A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In Proc. of UAI ’01, pages 437–444, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[9] P. Resnick and H. R. Varian. Recommender systems - introduction to the special section. Communication ACM, 40(3):56–58, 1997.

[10] A. Tveit. Peer-to-peer based recommendations for mobile commerce. In WMC ’01, pages 26–29, New York, NY, USA, 2001. ACM Press.

[11] D. J. Watts and S. H. Strogatz. Collective dynamics of ’small-world’ networks. Nature, 393(6684):440–442, June 1998.

[12] Y. Z. Wei, L. Moreau, and N. R. Jennings. A market-based approach to recommender systems.ACM Trans. Inf. Syst., 23(3):227–266, 2005

[13] Susan Gauch, Jeason Chaffee, and Alaxander Pretschner. 2003. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1, no. 3–4, pages 219–234.

[14] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics 28, no. 3, pages 245–288.

[15] Internet Movie Database. http://www.imdb.com.

Generic Trust-Based Resource Broker Architecture for Grid

Damandeep Kaur[1], Dr Jyotsna Sengupta[2]

[1]SPIC, Chandigarh,
[2]Department of Computer Science, Punjabi University, Patiala,
daman_811@yahoo.com,jyotsna.sengupta@gmail.com

Abstract**:** Grid resources and security issues go hand in hand in the success of any Grid application. The present research is moving towards achieving a secured resource management in Grid System, thereby allowing grid resource to enter the commercial area where the gird resource cannot be accessed through grid service without the assurance of a higher degree of trust relationship of resource provider. In this paper, we present a Generic Trust-based Resource Broker Architecture for Grid, along with various approaches which can be used in Trust Evaluation System to compute dynamic trust values which can be used to find degree of trust of grid resource providers**.**

Keywords**:** Trust, Reputation, Grid Resource Management, Grid Services

1. Introduction

Grid resources and security issues play a vital role in the success of any Grid application. Grid security research and development turns around better solutions to take care of the following requirements: Authentication, Secure Communication, Effective Security Policies, Authorization and access control where as RMS focuses mainly on handling the following a) geographical distribution of resources b) resource heterogeneity c) autonomously administered Grid Domains having their own resource policies and practices d) Grid domains using different access and cost models. A secured Grid Resource Management System allows it to enter the commercial area and without the assurance of a higher degree of trust relationship between consumer and provider, this cannot be achieved.

Trust and reputation mechanisms are used for large open systems. In general, *reputation* is the public's opinion about the character or standing (such as honesty, capability, reliability) of an entity, which could be a person, an agent, a product or a service. It is objective and represents a collective evaluation of a group of people/agents, while *trust* is personalized and subjective reflecting an individual's opinion. Trust can be transitive [1]. For example, Alice trusts her doctor and her doctor trusts an eye specialist. Then Alice can trust the eye specialist. The notions "trust" and "reputation" are closely related. Trust can be gained from a person/agent's own experiences with an entity or the reputation of the entity, while an entity's reputation relies on the aggregation of each individual person/agent's experiences with it. Trust and reputation are both used to evaluate an entity's trustworthiness. In this paper, in section 2, we present a Generic Trust-based Resource Broker Architecture for Grid and in Section 3, various approaches are discussed which can be used in Trust Evaluation System to compute dynamic trust values which can be used to find degree of trust.

2. Generic Trust-based Resource Broker Architecture for Grid

2.1 Layered Architecture

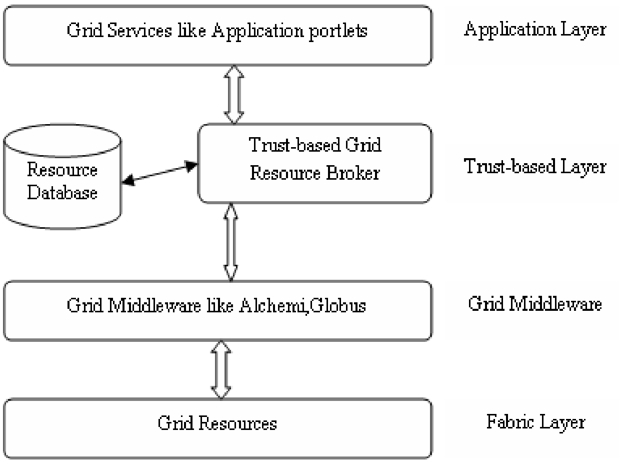

Grid Services like Application portlets
Application Layer
Resource Database
Trust-based Grid Resource Broker
Trust-based Layer
Grid Middleware like Alchemi,Globus
Grid Middleware
Grid Resources
Fabric Layer

Figure 1: Layered Architecture of Trust-based Grid

The layered architecture of our trust-based grid is shown in Figure 1.

i) Fabric Layer: This layer represents all the physical infrastructure of the Grid, including computers and the communication networks. It is made up of the actual resources that are part of the Grid, such as computers, storage systems, electronic data catalogues and even sensors such as telescopes or other instruments, which can be connected directly to the network.

ii) Middleware Layer: This layer refers to the grid middleware that incorporates necessary components for authentication, monitoring and discovery of grid resources, execution of job in grid resources, file transfer between grid resources.

iii) Trust Based Layer: This layer evaluates the trust value of all the grid resource providers. It computes overall trust value using any Trust Evaluation System and stores them in the database. This trust value is used to identify the most trusted resources for job execution. Suitable grid resources that match the job requirements are discovered and they are ranked on the basis of their trust value. The resource that has most trusted value is selected for grid services.

iv) Application layer: The highest layer of the structure is the application layer, which includes all different user applications (science, engineering, and business, financial), portals and development toolkits supporting the applications. This is the layer that users of the Grid will see and interact with.

2.2 Trust-Based Grid Resource Broker

The generic resource broker [9] consists of four main components, each having the basic functionality of providing standard interfaces to the rest of the application. The proposed Trust-based Grid Resource Broker shown in Figure 2 also has four components and each of these components is discussed next.

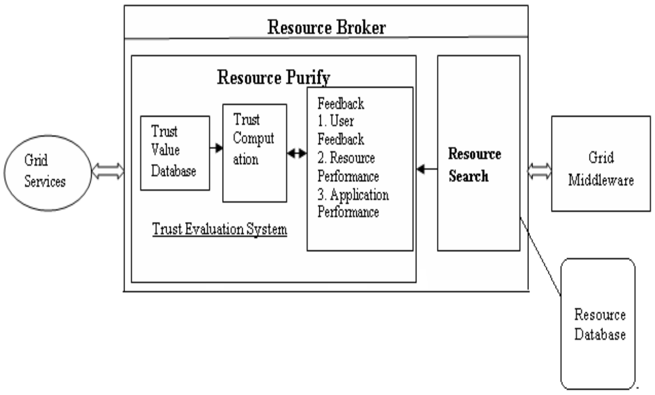


Figure 2: Trust-based Grid Resource Broker

1. Resource Search is a module responsible for fetching the list of all grid resources using any Grid Middleware tool like MDS, Gangalia and storing the list in the resource database. The Resource Search does not do any sort of data organization or processing, other than storing it in the resource database as it was received by the information source.

2 Resource Purify: It provides a standard interface where the implementation of the selection process can be easily adapted for various formats of user request. In the current implementation of the resource broker, the trust value of all grid resource providers is evaluated using various approaches discussed in next section and facilitates the selection of suitable resource for job execution based on the trust value.

It computes trust value of a resource provider based on any of the following three factors:-

- Infrastructure of the organization that provides a grid resource to the grid
- User FeedBack
- Performance metrics of the particular grid resource.

a) Feedback Factor
The feedback value is computed using three basic parameters:

1. User Feedback: This value is obtained using the user's feed back about a particular resource provider by prompting user to mention the level of satisfiability and willingness to recommend the resource to others. The two parameters namely the level of satisfiability and willingness to recommend are collectively called as user feedback parameters and they reflect the behavior of resource provider with user community.

2. Resource Performance /Application Performance: This value is obtained using various performance metrics of every resource providers or application and uses them in evaluating their trust. Various performance metrics like availability, number of success, number of failure, actual execution time, bandwidth, and latency can be considered.

b) Trust Evaluation System
This module can choose any of the various approaches discussed in next section, to compute the value of trust metrics received from underlying resources to calculate overall trust value and gathers the input from all the above modules and computes the overall trust of a resource provider and stores the value in trust value database, which acts as a small buffer. The trust value here represents the trustworthy of the resource provider at a given instant of time.

3. Various Trust-based Approaches

i) Trust Evaluation System can adopt various approaches which have already been applied in Grid

a) Quiz and Replication [15]: The basic idea is to insert indistinguishable quiz tasks with verifiable results known to the client within a package containing several normal tasks. The client can then accept or reject the normal tasks results based on the correctness of quiz results. By coupling Replication and Quiz, a client can potentially avoid malicious hosts and also reduce the overhead of verification and by adjusting the degree of result verification according to the trust value of a level of accuracy.

b) QoS[16]: With proposition that Grid resources and tasks are based on the trust QoS offset value, the most appropriate resources are allocated to specific service requests, which endeavors to achieve high QoS benefit value and allocate Grid resources on demand.

The trust relations between resources and tasks are classed into strong trust relation, weak trust relation and no trust relation. Since this algorithm aggregates tasks and resources based on matching offset and acquires QoS utility as much as possible in resource selection phase, meanwhile gets the smaller QoS matching offset and gains higher effective resource utility.

c) Managing Behavior of Resources [17]: basic idea of the approach is to view the interaction process between Grid participants similar to an industrial production process, and use statistical methods of quality assurance to discover deviations in the behavior of Grid participants in order to assess their behavior trust. Continuous Sampling Plan approach for managing the behavior trust of Grid participants is presented.

The aim of a continuous sampling plan (CSP) is to control the verification process depending on the verification results in such a way that the maximum of the average outgoing quality (AOQ) does not exceed a specified limit. AOQ can be defined as the fraction of "defective/non-conforming" entities which are not detected through the verification process with respect to the total number of processed entities.

d) D_S theory [18]: The paper proposes a method to detect the supply situation of Grid resources based on D-S theory which is monitored by trust function and trust lost function.

In addition, we propose the representation and updating mechanism of trust function and likelihood function, which calculate the nodes' trust through detecting the cost of receiving trust of nodes in Grid environments. What's more, we have proved the speculation trust function is sensitive and timely in simulating experiments.

In the Grid environment, how to deal with fault-tolerant of the unreliable Grid service and enhances the use factor of the entire Grid system is a future problem waiting for research.

e) Fuzzy-Logic [12]: This trust model combines first-hand (direct experience) and second-hand (reputation) information to allow peers to represent and reason with uncertainty regarding other peers' trustworthiness.

Fuzzy logic can help in handling the imprecise nature and uncertainty of trust. Linguistic labels are used to enable peers assign a trust level intuitively. Fuzzy trust model is flexible such that inference rules are used to weight first-hand and second-hand accordingly.

f) Self Protection [19]: This approach, intends to offer trust and reputation aware security for resource selection in grid computing. The Trust Factor (TF) value of each entity is determined from the self-protection capability and reputation weight age of that particular entity. Moreover the jobs are preferably assigned to the entities with higher TF values. The proposed approach has been found to cope with the ascending number of user jobs and grid entities. It aggregates several security related attributes for both self-protection capability and reputation into numerical values, which can be easily applied to calculate the Trust factor of grid entity. This scheme scales well with both number of jobs and number of Grid sites. This approach is quite effective in selecting secured entity for job execution from the available ones.

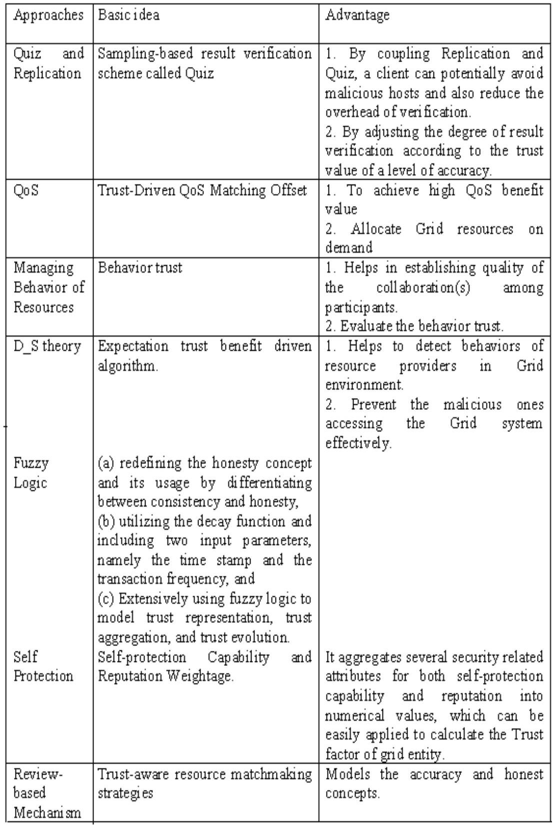
g) Review-based Mechanism [11]: This approach uses an *accuracy* concept to enable peer review-based mechanisms to function with imprecise trust metrics, the imprecision is introduced by peers evaluating the same situation differently. Simulation results show that the reputation-based trust model reaches an acceptable level of capability after a certain number of transactions. However, as the number of dishonest domains increase, the model becomes slow in reaching the acceptable level of capability.

To reduce the trust model's sensitivity to dishonest domains, we introduced an *honesty* concept to handle the situation where domains intentionally lie about other domains for their own benefit. Another feature of our model is the flexibility to weigh direct trust and reputation differently. Another significant advantage of our scheme is that our scheme does not depend on a majority opinion as previous schemes did. Therefore, our scheme can work even when majority of the recommenders are malicious.

Actually as the malicious number of recommenders increase, the recommenders providing recommendations to a query reduces. The number of recommenders also provides another measure of trust on the overall system because all the recommenders are considered honest.

ii) Some Open Issues

1. To find efficient and scalable mechanisms for generating quizzes.
2. How to deal with fault-tolerant of unreliable Grid service
3. Focus on trust-driven DAG task scheduling
4. Accuracy and demand for a variety of reputation systems and verification schemes should be conducted.
5. A cooperative scheme to detect and estimate the fraction of malicious nodes behavior to make Accuracy on Demand more precise.



Approaches	Basic idea	Advantage
Quiz and Replication	Sampling-based result verification scheme called Quiz	1. By coupling Replication and Quiz, a client can potentially avoid malicious hosts and also reduce the overhead of verification. 2. By adjusting the degree of result verification according to the trust value of a level of accuracy.
QoS	Trust-Driven QoS Matching Offset	1. To achieve high QoS benefit value 2. Allocate Grid resources on demand
Managing Behavior of Resources	Behavior trust	1. Helps in establishing quality of the collaboration(s) among participants. 2. Evaluate the behavior trust.
D_S theory	Expectation trust benefit driven algorithm.	1. Helps to detect behaviors of resource providers in Grid environment. 2. Prevent the malicious ones accessing the Grid system effectively.
Fuzzy Logic	(a) redefining the honesty concept and its usage by differentiating between consistency and honesty, (b) utilizing the decay function and including two input parameters, namely the time stamp and the transaction frequency, and (c) Extensively using fuzzy logic to model trust representation, trust aggregation, and trust evolution.	
Self Protection	Self-protection Capability and Reputation Weightage.	It aggregates several security related attributes for both self-protection capability and reputation into numerical values, which can be easily applied to calculate the Trust factor of grid entity.
Review-based Mechanism	Trust-aware resource matchmaking strategies	Models the accuracy and honest concepts.

4. Conclusion and Future Work

This paper introduces a generic trust-based resource broker that bridges the gap between a user's requirement and secured grid services on the basis of trust values, along with various approaches for Trust Evaluation System. Finally we have also tried to address few open issues in regard to the future direction of the research.

References

[1] A. Jøsang, L. Gray and M. Kinateder. Simplification and Analysis of Transitive Trust Networks 4(2) 2006,pp.139-161 . Web Intelligence and Agent Systems Journal. 2006

[2] L.-H. Vu, M. Hauswirth, K. Aberer, QoS-based service selection and ranking with trust and reputation management, Proceedings of OTM'05, R. Meersman and Z. Tari (Eds.), LNCS 3760, p.p. 466-483, 2005.

[3] I. Foster, C. Kesselman, and S. Tuecke, “The anatomy of the Grid: Enabling scalable virtual organizations." Int. J. Supercomputing, vol. 15, no. 3, pp. 200-222, 2001.

[4] Foster, I. and Kesselman, C. Computational Grids. Foster, I. and Kesselman, C. eds. The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 1999, 2-48.

[4] I. Foster. Globus toolkit version 4: Software for service oriented systems. In Proc. of the IFIP International Conference on Network and Parallel Computing, 2005.

[5] R. Buyya and S. Venugopal, The Gridbus Toolkit for Service Oriented Grid and Utility Computing: An Overview and Status Report, Proceedings of the First IEEE International Workshop on Grid Economics and Business Models (GECON), 2004.

[6] V.Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S.Meder, L. Pearlman and S. Tuecke, “Security for Grid Services”, in Proceedings of the HPDC-12, 2003.

[7] Foster, I., Kesselman, C., Tsudik, G. and Tuecke, S. A Security Architecture for Computational Grids. ACM Conference on Computers and Security, 1998, pp: 83-91.

[8] Farag Azzedin, Muthucumaru Maheswaran, "Towards Trust-Aware Resource Management in Grid Computing Systems," ccgrid, p. 452, 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02), 2002.

[9]Enis Afgan, Vijay Velusamy, Purushotham V. Bangalore, Grid Resource Broker Using Application Benchmarking, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Advances in Grid Computing - EGC 2005, Volume 3470/2005 ,Pages 691-701.

[10]. F. Azzedin and M. Maheswaran, “A Trust Brokering System and Its Application to Resource Management in Public- Resource Grids”, in Proceedings of IPDPS 2004.

[11] S. Song, K. Hwang and M. Macwan, “Fuzzy Trust Integration for Security Enforcement in Grid Computing”, in Proceedings of IFIP International Conf. on Network and Parallel Computing, (NPC-2004), Wuhan, China, October 18–20, 2004, pp. 9–21.

[12] L. Xiong and L. Liu, “PeerTrust: Supporting Reputationbased Trust to P2P E-Communities”, IEEE Trans. Knowledge and Data Engineering, July 2004, pp. 843–857.

[13] Chunqi Tian, Shihong Zou, Wendong Wang, Shiduan Cheng,An Efficient Attack-Resistant Trust Model for P2P Networks, IJCSNS, Vol. 6 No. 11 pp. 251-258, 2006.

[14] Baolin Ma, Jizhou Sun, Ce Yu, Reputation-based Trust Model in Grid Security System, Journal of Communication and Computer, Volume 3, No.8 (Serial No.21), 2006.

[15]. S. Zhao and V. Lo, “Result Verification and Trust-based Scheduling in Open Peer-to-Peer Cycle Sharing Systems,” in IEEE Fifth International Conference on Peer-to-Peer Systems, Sept. 2005.

[16] Ran Li, Jiong Yu: QoS Matching Offset Algorithm Based on Trust-Driven for Computing Grid International Conference on Computer Science and Software Engineering, CSSE 2008, Volume 3: Grid Computing / Distributed and Parallel Computing / Information Security, December 12-14, 2008, Wuhan, China. 170-173.

[17] Elvis Papalilo and Bernd Freisleben, Managing Behaviour Trust in Grid Computing Environments, Journal of Information Assurance and Security, Volume 3, Issue 1, March 2008, page 27-38.

[18] Yongsheng Hao, Guanfeng Liu, Yuebin Xu and Junyan Wang, A New Expectation Trust Benefit Driven Algorithm for Grid Environments, International Journal of Hybrid Information Technology Vol. 2, No. 1, January, 2009.

[19] V.Vijayakumar and Dr. R.S.D. Wahida Banu, “Security for Resource Selection in Grid Computing Based On Trust and Reputation Responsiveness," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.Page 107-118.

Author Biographies

Damandeep Kaur, 1980 holds M.E (Software Engineering) degree form Thapar University, Patiala. She is currently pursuing PhD from Punjabi University Patiala and working as Senior Software Engineer in Department of Excise & Taxation, Chandigarh.. Her area of interests includes Parallel and Distributed Computing, Grid Computing, Resource Management, Resource Discovery and Peer-to-Peer Networks.

Jyotsna Sengupta holds PhD degree from Thapar University, Patiala. She is currently H.o.D & reader in Department of Computer Science, Punjabi University, Patiala. Her area of interests includes Parallel and Distributed Computing, Grid Computing, Adhoc Networks.

Effective Classification Technique Enhanced Using Genetic Algorithm: For Data Mining Disease in the Incumbents to the Health Centre

Manaswini Pradhan[1] and Dr. Ranjit Kumar Sahu[2]

[1]Lecturer, P.G. Department of Information and Communication Technology,
Fakir Mohan University, Orissa, India
E-mail: ms.manaswini.pradhan@gmail.com

[2] Consultant, Plastic, Cosmetic and Laser Surgery, Mumbai, India
E-mail: drsahurk@yahoo.com

Abstract*: The diagnosis of disease is a vital and intricate job in the incumbents to the health centre. The proposed method combines the learning algorithm of BP neural network with genetic algorithm to train BP network and optimize the weight values of the network in a global scale. This method is featured as global optimization, high accuracy and fast convergence. The data-mining model based on genetic neural network is selected as the enhanced classifier and has been widely applied to the procedure of data mining on case information of incumbents in the reception counter of a health centre. It achieves an excellent effect for assisting health professionals to solve cases and make good decisions. In this paper, the principles and methods of this data-mining model are described in details. A real case of its application is also presented which predicts the disease in the incumbents to the health centre. From this case we can draw a conclusion that the data-mining model we have chosen is scientific, efficient, robust and practicable.*

Keywords: *Data Mining, Data Warehouse, BP Neural Network, Genetic Algorithm, Data Cleaning, Case Analysis*

1. Introduction

Data mining is the method of extraction of information and knowledge that are hided in data, unknown by people and potentially useful from a huge amount of data with multiple characteristics that is incomplete, containing noise, fuzzy and random. As a kind of cross-discipline field that combines multiple disciplines including database technology, artificial intelligence, neural networks, statistics, knowledge acquirement and information extraction, nowadays data mining has become one of the most important research direction in the international realms of information-based decision making. Analyzing and comprehending data from different aspects, people use data mining methods to dig out useful knowledge and hidden information of prediction from a large amount of data that are stored in database and data warehouse. The methods include association rules, classification knowledge, clustering analysis, tendency and deviation analysis as well as similarity analysis. By finding valuable information from the analysis results, people can use the information to guide their business actions and administration actions, or assist their scientific researches. All of these provide new opportunities and challenges to the development of all kinds of fields related to data processing.

Data mining is applied to the procedure of data analyzing, processing, decision making and data warehouse. Data mining technologies assist in many social departments to make scientific and reasonable decisions. This has major contribution for the development of our society and economy. Data mining can be applied to various different realms. For instance, many sale departments use data mining technology to determine the distribution and the geographical position of the sale network, the purchase and stock quantities of every kind of goods, in order to find out the potential customer groups and adjust the strategies for sale. In insurance companies, stock companies, banks and credit card companies, people apply data mining technology to detect the deceptive actions of customers to reduce the commercial deceptions. Data mining has been also widely applied to medical treatment and genetic engineering and many other fields. In recent years, with the acceleration of the step of information construction in police departments and with the increment of its development level, data mining technology has also been applied to the health departments especially in the health centre to improve the hospital treatment. This paper mainly discusses the principle and the practical application of genetic neural network based data mining model in disease analysis of patients.

Data classification is a classical problem extensively studied by statisticians and machine learning researchers. It is an important problem in variety of engineering and scientific disciplines such as biology, psychology, medicines, marketing, computer vision, and artificial intelligence A.K. Jain *et al*(2000). The goal of the data classification is to classify objects into a number of categories or classes. There have been wide ranges of machine learning and statistical methods for solving classification problems. Different parametric and non-parametric classification algorithms have been studied R.O. Duda *et al*.(1973), Breitman. L *et. al*(1984), Buntine, W.L *et al*(1992),Cover, T. M. *et al*.(1967), Hanson R. *et al*.(1993), Michie,D, *et al*(1994), Richard, M.D *et al*.(1991) and Tsoi, A.C. *et al*(1991). Some of the algorithms are well suited for linearly separable problems. Non-linear separable problems have been solved by neural networks dealt by C. Bishop (1995), support vector machines V.N. Vapnik *et al*. (1971) etc.

Neural networks (NNs) are increasing in popularity due their ability to approximate unknown functions to any degree of desired accuracy, as demonstrated by Funahashi (1989) and Hornik *et al*. (1989). In addition, NNs can also do this without making any unnecessary assumptions about the distributions of the data. This makes it convenient for researchers, as it can include any input variables that they feel could possibly contribute to the NN model. Although, it is likely that irrelevant variables are introduced to the model, the NN is expected to learn sufficiently to ignore these variables during the training process. It does this by finding weights associated to these irrelevant variables that when plugged into the NN would generate values that simply zero each other out, thereby having no effect on the final output prediction.

Although this works fine for training data, when applied to observations that it has not seen (out-of-sample or testing data), these weights are going to generate values that are unlikely to zero each other out, causing additional error in the prediction. If, on the other hand, the research could identify the irrelevant variables, these variables could be excluded from the NN model and eliminate the possibility of introducing additional error when applied to out of-sample data. Although, it is convenient for researchers to be able to include all available input variables into the model to extract a good solution, it also has the detrimental effect of making the NN a ‘black box’ where they throw everything into the model but do not know why or how the network produces its output. Additional information about the problem can be obtained by identifying those inputs that are actually contributing to the prediction. For example, we could train a NN that predicts the disease in the incumbent to the health centre. As inputs, we could include patient’s information such as gender, age, education of the incumbent, disease history, salary level and bad habits etc. A NN model that can accurately predict this outcome as well as indicating the relevant inputs to the model would be very beneficial in identifying disease. By using the proposed algorithm to determine those variables that are relevant to prediction, additional information about the problem can be learned.

The next section includes a background of literature on back propagation and the genetic algorithm and describes the problem. The third section describes the Data mining model of enhanced supervised classifier. The fourth section describes the GA method used in this study, which includes the base algorithm. The fifth section describes the problem, how the data were generated and outlines how the GA determines the number of hidden nodes (architecture) and the training process. Reports of the results of the application of Data mining in the Health Centre to diagnose the disease in the incumbent are performed. The last section concludes with final suggestions.

2. Background Literature

Since the majority of NN research is conducted using gradient search techniques, such as back propagation, which require differentiability of the objective function, the ability for researchers to identify relevant variables, beyond trial and error, is eliminated. In this paper, a modified genetic algorithm is used for training a NN, which does not require differentiability of the objective function that will correctly distinguish relevant from irrelevant variables and simultaneously search for a global solution.

2.1 Backpropagation

Back propagation (BP) is currently the most widely used search technique for training NNs. BP’s original development is generally credited to Werbos (1993), Parker (1985) and LeCun (1986) and was popularized by Rumelhart et al. (1986a,b). Although many limitations to this algorithm have been shown in the literature (Archer and Wang, 1993; Hsiung *et al*., 1990; Kawabata, 1991; Lenard *et al*., 1995;Madey and Denton, 1988; Masson and Wang, 1990; Rumelhart et al., 1986a; Subramanian and Hung, 1990; Vitthal *et al*., 1995; Wang, 1995;Watrous, 1987;White, 1987), its popularity continues because of many successes. An additional limitation to BP, which this paper deals with, is its inability to identify relevant variables in the NN model. This inability stems from the gradient nature of BP, which requires the objective function (usually the sum of squared errors (SSEs)) to be differentiable. This requirement prevents any attempt to identify weights in the models that are unnecessary, beyond pruning of the network. Pruning the network is simply eliminating connections that have basically no effect on the error term. This can be done by trial and error, saliency of weights, and node pruning (Bishop, 1995). Also, there have been approaches to network construction, such as Cascade Correlation (Fahlman and Lebiere, 1990), which attempts to build parsimonious network architectures. A better approach might be to use an alternative algorithm, such as the GA, that is not dependent on derivatives to modify the objective function to penalize for unneeded weights in the solution. By doing so, the GA can search for an optimal solution that can identify those needed weights and corresponding relevant variables.

2.2 The Genetic Algorithm

The GA is a global search procedure that searches from one population to another for solutions, focusing on the area of the best solution as far as practicable, while continuously sampling the total parameter space. Unlike back propagation, the GA starts at multiple random points (initial population) when searching for a solution. Each solution is then evaluated based on the objective function. Once this has been done, solutions are then selected for the second generation based on how well they perform. Once the second generation is drawn, they are randomly paired and the crossover operation is performed. This operation keeps all the weights that were included in the previous generation but allows for them to be rearranged. This way, if the weights are good, they still exist in the population. The next operation is mutation, which can randomly replace any one of the weights in the population in order to find a solution so as to escape local minima. Once this is complete, the generation is ready for evaluation and the process continues until the best solution is found. The GA works well searching globally because it searches from many points at once and is not hindered by only searching in a downhill fashion like gradient techniques. Schaffer *et al*. (1992) found more than 250 references in the literature for research pertaining to the combination of genetic algorithms and NNs. In this research, the GA has been used for finding optimal NN architectures and as an alternative to BP for training. This paper combines these two uses in order to simultaneously search for a global solution and a parsimonious NN architecture. Schaffer (1994) found that most of the research using the GA as an alternative training algorithm has not been competitive with the best gradient learning methods. However, Sexton et al. (1998) found that the problem with this research is in the implementation of the GA and not its inability to perform the task. The majority of past implementations of the GA encode each candidate solution of weights into binary strings. This approach works well for optimization of problems with only a few variables but for neural networks with a large number of weights, binary encoding results in extremely long strings. Consequently, the patterns that are essential to the GA's effectiveness are virtually impossible to maintain with the standard GA operators such as crossover and mutation. It has been shown by Davis (1991) and Michalewicz (1992) that this type of encoding is not necessary or beneficial. A more effective approach is to allow the GA to operate over real valued parameters (Sexton, Dorsey, and Johnson, 1998). The alternative approach described in Sexton *et al*. (1998) also successfully outperformed back propagation on a variety of problems. This line of research in based on the algorithm developed by Dorsey and Mayer (1995) and Dorsey *et al*. (1994). The GA and its modifications used in this study follows in the next section. Since the GA is not dependent on derivatives, a penalty value can be added to the objective function that allows us to search not only for the optimal solution but also for one that identifies relevant inputs to the model.

2.3 Data Mining In Reception Counter of a Health Centre

2.3.1 The Meaning of Data Mining in Reception Counter of a Health Centre

Every day in the reception counter of any health centre, patients arrive with different diseases. These large numbers of disease cases are received with various approaches. The information has been input into database to form a large amount of disease case information. These disease case information has been archived annually and periodically to form a plenty of historical case resources. By inducing and analyzing these historical cases, physician and people can get some experiences and learn some lessons that can help them to solve cases and make decisions in the future for getting better and improved health facilities. Therefore, in order to assist health departments to solve cases rapidly and make decisions efficiently, we should synthesize and organize these historical data, use proper data mining models to discover the potential and useful knowledge behind the data, and then predict and analyze the important factors in the data including the rate of disease, the constitution of disease population, the disease age structure, the area distribution of disease, the developing tendency of disease, the means and approaches of disease, the hidden areas of disease and so on. At present all of these have become urgent tasks that need our health centres to accomplish in the procedure of data processing.

2.3.2 Steps of Data Mining

The data mining steps in the health centers mainly include two steps:

(1) Filtering, selecting, cleaning and synthesizing the archived historical case information, and then performing transformation, if necessary, and finally, loading data into data warehouse after the above processing.

(2) Choosing appropriate models and algorithms of data mining to find out the potential knowledge in data. By a number of analysis and comparison among various data mining models, we select the back propagation (BP) error neural network as the general-purpose calculation model in our data mining. We train the neural network with a supervised learning method and combine BP algorithm with genetic algorithm to optimize the values of weights. Further, we apply the trained model to the prediction, classification and rule extraction of the case information.

3. Data Mining Model of Enhanced Supervised Classifier

3.1 General Methods of Data Mining

Now-a-days data mining methods include statistical method, association discovery, clustering analysis, classification and regression, OLAP(On Line Analytical Processing), query tool, EIS(Executive Information System), neural network, genetic algorithm and so on. Because of its high sustenance to noise data, good ability of generalization, high accuracy and low error rate, neural network model possesses great advantages among data mining methods. Hence, it has become a popular tool in data mining.

3.2 Data Mining Model of BP Neural Network

BP neural network is a kind of feed forward network that is now in most common use. Generally it has a multi-layer structure that consists of at least three layers including one input layer, one output layer and one or more hidden layers. There are full connections between neurons in adjacent layers and no connection between neurons in the same layer. Based on a set of training samples and a set of testing data, BP neural network trains its neurons and complete the procedure of learning. The application of BP algorithm is suitable for data mining environment in which it is impossible to solve problems using ordinary methods. Therefore, we need the use of complex function of several variables to complete non-linear calculation to accomplish the semi-structural and non-structural decision-making supporting procedure. So in the procedure of data mining in the reception counter of a health centre, we choose the BP neural network model.
The basic structure of BP neural network is as follows:

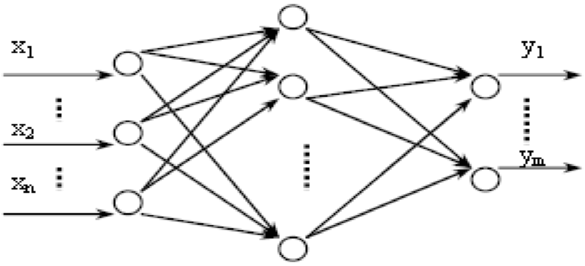


Fig. 1 The Structure of BP Neural Network

The learning procedure of neural network can be divided into two phases:

(1) The first one is a forward propagation phase in which a specified input pattern has been passed through the network from input layer through hidden layers to the output layer and becomes an output pattern.

(2) The second one is an error back propagation phase. In this phase, BP algorithm compares the real output and the expected output to calculate the error values. After that, it propagates the error values from output layer through hidden layer to input layer in the opposite direction. The connection weights will be altered during this phase.

These two phases proceed repeatedly and alternately to complete the memory training of network until it tends to convergence and the global error tends to minimum.

3.3 Learning Algorithm of Proposed Enhanced Classifier

In practical application of data mining, we use the three-layer BP neural network model that includes a single hidden layer and select differentiable Sigmoid function as its activation function. The function is defined as formula (1):

$$f(x)=1/(1+e^{-x}) \qquad (1)$$

The learning algorithm of BP neural network is described as follows:

(1) Setting the initial weight values *W (0)*: Generally we generate random nonzero floating numbers in [0, 1] as the initial weight values.

(2) Choosing certain numbers of pairs of input and output samples and calculate the outputs of network. The input samples are $X_{s=}(x_{1s,}x_{2s,\dots,}x_{ns})$. The output samples are $t_s=(t_{1s,}t_{2s,\dots,}t_{ms})$, s=1,2,…L. L is the number of input samples. When the input sample is the sth sample the output of the ith neuron is y_{is} :

$$y_{is}(t)=f\ (\sum_j w_{ij}(t)x_{js}) \qquad (2)$$

(3) Calculating the global error of network. When the input sample is the sample is the error of network. The calculating formula of is the sth *sample* E_s is the error of network. The calculating formula of E_s is:

$$E_s(t)= \frac{1}{2\sum_k (t_{ks}-y_{ks}(t))^2}$$

$$= \frac{1}{2\sum_k e_{ks}{}^2(t)} \qquad (3)$$

where *k* represents the *kth* neuron of output layer. y $_{ks}$(t) the output of network when input sample is the sample *Sth* sample and the weight values has been adjusted *t* times. After training network *t* times based on all of the *L* groups of samples, the global error of all of these samples is:

$$G(t)=\sum_s E_s(t) \qquad (4)$$

(4) Determining if the algorithm ends.

$$G(t)\le\varepsilon \qquad (5)$$

When the condition of formula (5) is satisfied the algorithm ends. ε is the limit value of error that is specified beforehand. ε >0.

(5) Calculating the error of back propagation and adjusting the weights. The gradient descent algorithm has been used to calculate the adjustment values of weights. The calculating formula is as follows:

$$W_{ij}(t+1)=W_{ij}(t)-\ \eta\ \frac{\partial G(t)}{\partial w_{ij}(t)}$$

$$=W_{ij}(t)\ -\ \eta \sum_s \frac{\partial E_s(t)}{\partial w_{ij}(t)} \qquad (6)$$

where η is learning rate of network and also the step of weight adjustment.

3.4 Difficulties of the BP Network and the appropriate Solution

Because we use the gradient descent algorithm to calculate the values of weights, BP neural network still encounters problems such as local minimum, slow convergence speed and convergence instability in its training procedure. We combine two methods to solve these problems. One solution is to improve the BP network algorithm. By adding steep factor or acceleration factor in activation function, the speed of convergence can be accelerated. In addition, by compressing the weight values when they are too large, the network paralysis can be avoided. The improved activation function is defined with formula (7):

$$f_{a,b,\lambda}(x)=\frac{1}{1+e^{(x-b)/\lambda}}+\alpha \qquad (7)$$

where is α deviation parameter, *b* is a position parameter and *λ* is the steep factor.

Another solution to this is that the Genetic algorithm is a concurrence global search algorithm. Because of its excellent performance in global optimization, we can combine the genetic algorithm with BP network to optimize the connection weights of BP network. And finally we can use the BP algorithm for accurate prediction or classification.

4. Genetic Algorithm to Enhance BP Neural Network

4.1 The Principle of Genetic Algorithm

Genetic algorithm is a kind of search and optimization model built by simulating the lengthy evolution period of heredity selection and natural elimination of biological colony. It is an algorithm of global probability search. It doesn't depend on gradient data and needn't the differentiability of the function that will be solved and only need the function can be solved under the condition of constraint. Genetic algorithm has powerful ability of macro scope search and is suitable for global optimization. So by using genetic algorithm to optimize the weights of BP neural network we can eliminate the problems of BP network and enhance the generalization performance of the network.

The individuals in genetic space are chromosomes. The basic constitution factors are genes. The position of gene in individual is called locus. A set of individuals constructs a population. The fitness represents the evaluation of adaptability of individual to environment.

The elementary operation of genetic algorithm consists of three operands: selection, crossover and mutation. Selection is also called copy or reproduction. By calculating the fitness f_i of individuals, we select high quality individuals with high fitness, copy them to the new population and eliminate the individual with low fitness to generate the new population. Generally used strategies of selection include roulette wheel selection, expectation value selection, paired competition selection and retaining high quality individual selection. Crossover puts individuals in population after selection into match pool and randomly makes individuals in pairs to form parent generation. Then according to crossover probability and the specified method of crossover, it exchanges part of the genes of individuals that is in pairs to form new pairs of child generation and finally to generate new individuals. Generally used methods of crossover are one point crossover, multi point crossover and average crossover. According to specified mutation rate, mutation substitutes genes with their opposite genes in some loci to generate new individuals.

4.2 The Calculating Steps of Genetic Algorithm

The methods and steps of utilizing genetic algorithm to optimize the weights of BP network are described as follows:

(1) First, *k* groups of weights are given at random and assigned to *k* sets of BP networks. By training the networks, *k* groups of new weights has been calculated and adjusted. They constitute the original solution space.

(2) Using real number coding method these weights are coded to decimals and used as chromosomes. *k* groups of chromosomes comprise a population. So the original solution space has been mapped to search space of genetic algorithm. The length of gene string after coding is L=m×h+h×n . Where *m* is the number of neutrons in input layer, η is the number of neurons in hidden layer and *n* is the number of neurons in output layer.

(3) Using minimum optimization method the fitness function can be determined. The formula of fitness function is as follows:

$$f=\frac{1}{2G}=\frac{1}{\sum_{i=1}^{s}\sum_{j=1}^{m}(t_{ij}-y_{ij})^2} \qquad (8)$$

where is *S* the total number of samples, *m* is the number of neurons in output layer, *G* is the global error of all of *S* numbers of samples and y_{ij} is the output of network.

(4) The weights are optimized using genetic algorithm. We calculate the fitness and perform the selection with method of roulette wheel selection. After that, we copy the individuals with high fitness into next generation of the population. The next step is crossover. We crossover the individuals after selection with probability $P_{c.}$ Because we use real number coding method to code weights into decimals, the algorithm of crossover should be altered. If the crossover is performed between the i_{th} individual and the $(i+1)^{th}$ individual, the operand is as follows:

$$x_i^{i+1}=c_{i_i}*x_i^t+(1-c_i)*x_{i+1}^t$$

$$x_{i+1}^{i+1} = (1 - c_{i_i}) * x_i^t + c_i * x_{i+1}^t \qquad (9)$$

where is x_i^t, x_{i+1}^t a pair of individuals before crossover, x_i^{t+1}, x_{i+1}^{i+1} is a pair of individuals after crossover. c_i is a random datum of uniform distribution in [0,1] . With probability P_m, we mutate the individuals after crossover. If we mutate the ith individuals, the operand is

$$x_i^{i+1} = x_i^t + c_i \qquad (10)$$

where x_i^t is an individual before mutation, x_i^{t+1} is an individual after mutation, c_i is a random datum of uniform distribution in [$u_{\min} - \delta_1 - x_i^t, u_{\max} + \delta_2 + x_i^t$]. After once of these operations, a new population is generated. By repeating the procedure of selection, crossover and mutation, the weight combination is adjusted close enough to the most optimized weight combination.

(5) Finally, through the BP networks the weights can be adjusted delicately. Till now, the whole procedure of optimization ends.

With respect to every kind of prediction and analysis problems in the course of data mining, we extract proper sets of training samples and testing data, train mature neural network models with above-mentioned methods and apply the models to the future case analysis and prediction.

5. Application of the Data mining in the Health Centre to Diagnose the Disease

Finally, we give a real application of data mining in the reception counter of health centre as example. In this example we analyze patient's gender, age, educational degree, history of disease, chronic/acute, personal features, social relations and economical incomes. We find that to some extent these factors affect patient's social actions and habits that may lead patient to suffer a disease. Using these factors as input variables, a genetic neural network can be utilized to predict the present disease possibility of the patients..

5.1 Clean the Data in Database

In the first step, we fill up the missing data, smooth the noise data in database and solve the problems of same name for different meaning and different name for same meaning. And then, we load related data into data warehouse.

5.2 Select Training Samples of BP Networks

As in case archive databases, the case information are arranged in order of time, representative data can be obtained by random sampling. So we select samples by random sampling. To obtain the training sample set of BP networks, we select 5000 records from data warehouse. In addition, we extract other 2000 records as the testing sample set.

5.3 Normalize Samples

The most important input variables of BP network include gender, age, education degree, disease history, salary level and bad habits. The output of samples is the status (Yes or No) of whether these people suffer a disease at present. The output of BP network is the probability of people's present status (Percentage) of disease. Table 1 gives a list of first 10 samples of the total 5000 training samples.

By normalizing above input and output variables, the range of values of these variables has been mapped to the range of [0, 1]. The mapping relationship is given as follows:

5.3.1 Gender

Male: 1.0;
Female: 0.0

5.3.2 Age

0: 0.00;
1: 0.01;
2: 0.02; ⋯;
100 and above: 1.0

No.	Sex	Age	Education Degree	Disease History	Salary Level	Bad Habits	Present Status of disease
1	M	25	Secondary school	Yes	3000--8000	No	Yes
2	M	32	Secondary school	No	1--3000	Yes	Yes
3	M	40	Primary School	Yes	3000--8000	Yes	Yes
4	F	30	Primary School	No	50000--80000	No	No
5	F	27	Secondary School	Yes	3000--8000	Yes	Yes
6	M	28	University	No	15000-30000	No	No
7	M	50	Junior University	No	8000--15000	No	No
8	M	38	Post-graduate	No	50000-80000	No	No
9	M	70	Primary School	No	1--3000	Yes	No
10	F	35	High School	No	3000--800 0	No	No

Table 1: Values of Input Variables

5.3.3 Education Degree

Illiterate: 0.0;
Graduate of Primary School: 0.125;
Graduate of Secondary School: 0.25;
Graduate of High School: 0.375;
Graduate of Junior University: 0.5;
Graduate of University: 0.625;
Postgraduate: 0.75;
Doctor: 0.875;
Post doctor: 1.0

5.3.4 Disease History

Yes: 1.0;
No: 0.0

5.3.5 Salary Level

None: 0.0;
Below 3000 Rupees: 0.125;
3000—8000 Rupees: 0.25;
8,000—15,000 Rupees: 0.375;
15,000—30,000 Rupees: 0.5;
30,000—50,000 Rupees: 0.625;
50,000—80,000 Rupees: 0.75;
80,000—1, 50,000 Rupees: 0.875;
1, 50,000 Rupees and above: 1.0

5.3.6 Bad Habits

Yes: 1.0; No: 0.0

5.3.7 Present Status of Disease

Yes: 1.0; No: 0.0

Table 1 gives the value list of first 10 samples of the total 5000 training samples after normalization.

No.	Sex	Age	Education Degree	Disease History	Salary Level	Bad Habits	Present Status of Disease
1	1	0.25	0.25	0	0.25	0	1
2	1	0.32	0.25	1	0.125	1	1
3	1	0.40	0.125	1	0.25	1	1
4	0	0.45	0.125	0	0.75	0	0
5	0	0.27	0.25	1	0.25	1	1
6	1	0.28	0.625	0	0.5	0	0
7	1	0.50	0.5	0	0.375	0	0
8	1	0.38	0.75	0	0.75	0	0
9	1	0.70	0.125	0	0.125	1	0
10	0	0.35	0.375	0	0.25	0	0

Table 2 Normalized Values of Input Variables

5.4 Build BP Neural Networks and Begin to Train

As it is observed, including above 6 important variables the total number of input variables is 10, we determined that the number of neurons in input layer is 10 and the number of neurons in output layer is 1. According to our experience and conforming to the principle of simplifying the network structure, we set the number of neutrons in hidden layer to 16. With above parameters we build 10 BP networks that have same structure. Then we generate 10 sets of small random numbers as initial weights of these networks and use the extracted 5000 samples as input and output samples of these networks. After that, we utilize BP algorithm to train the networks and get 10 sets of trained weights. The training times are 8000. After training we test the networks with our testing sample set. The generalization ability of our first network is shown as follows:

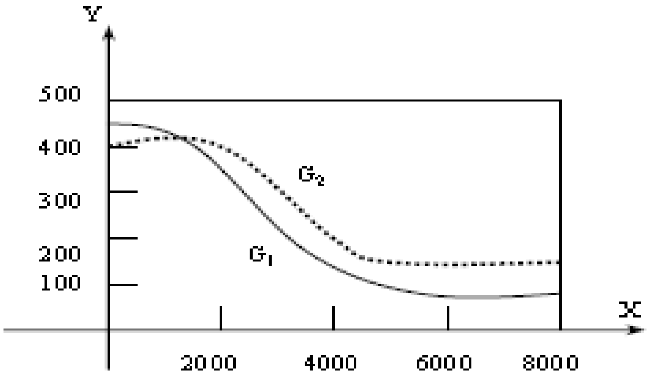


Fig. 2 The Generalization Ability of BP Network

where X is times of training, Y is the value of error, G_1 is global error of training sample set and G_2 is global error of testing sample set.

5.5 Utilize Genetic Algorithm to Optimize the Weight Values

We code the 10 sets of trained weights by real number coding method and use the weights after coding as chromosomes. 10 groups of chromosomes consist of a population. Then we optimize these weights using genetic algorithm until the weights, after decoding, are adjusted close enough to the most optimized weight combination.

5.6 Use the Optimized Weights to Train the BP Network Again

Finally, we use one of the BP network to adjust the optimized weights delicately. The training times for this adjustment are 4000. As a result, the generalization ability of the network is shown below:

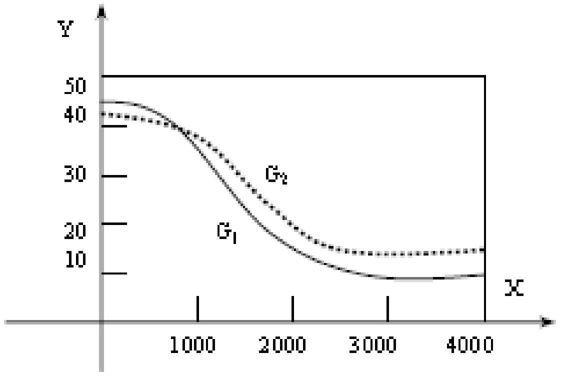


Fig. 3 The Generalization Ability of BP Network

5.7 Apply the Trained Network to Prediction and Analysis

We use the finally adjusted weights as the running weights of BP network to predict the probability of disease that people may suffer at present. The probability is the output of the BP network and is a float point number representing the occurrence probability of events. The prediction result by this process is highly accurate. In real terms of the disease diagnose of health centre, this prediction result can be used to guide the monitoring and tracing against the former attack of the disease to a person. At the same time, it can assist the lock and confirmation of suspects in case detection. Hence, it is highly useful and fool-proof method for case solving and decision-making.

6. Conclusion

GA was found to be an appropriate alternative to BP for training neural networks that not only finds better solutions with a parsimonious structure but can also identify relevant input variables in the data set. By using the GA in this manner, researchers can now determine those inputs that contribute to estimation of the underlying function. This can help with analysis of the problem, improved generalization, and network structure reduction. These results have demonstrated that a NN can be more than just a ‘black box’. A complex chaotic time series problem as well as real-world problems could be solved that outperformed traditional NN training techniques as well as discovering relevant input variables in the model. Based on these results, future research is warranted for additional experiments and comparisons using the GA for NN training. BP neural network that has been applied to data mining possesses characteristics of high ability of memory, high adaptability, accurate knowledge discovery, none restriction to the quantity of data and fast speed of calculation. Based on using genetic algorithm to optimize the BP network can effectively avoid the problem of local minimum. Therefore, enhanced supervised classifier which is the proposed data-mining model has many advantages over other data mining models. In the real practice of data mining in the disease diagnosis in health centre the advantages have been fully embodied. This method has its own usefulness and is an effective prediction system to detect any type of diseases and at the same time has its beneficial effect upon the society.

References

1. Aoying Zhou, 2005. A Genetic-Algorithm-Based Neural Network Approach for Short-Term Traffic Flow Forecasting. Advances in Neural Networks, 3498, pp. 965-969.

2. Archer NP, Wang S. 1993. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. Decision Sciences 24(1): 60–75.

3. Berson Alex, Smith Stephen J. Data Warehousing, Data Mining, & OLAP. McGraw-Hill Book Co, 1999

4. Bishop CM. 1995. Neural Networks for Pattern Recognition. Clarendon Press: Oxford.

5. Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.

6. D.E.Goldberg, 1989. Genetic Algorithms in Search, Optimization and Machine. Leaning, Addison-Wesley.

7. D.E.Goldberg, 1992. Genetic Algorithms: A Bibliography, IlliGAL Technical Report , 920008.

8. David Hard, 2003. Principles of Data Mining. Machine Industry Publisher, Beijing.

9. Davis L (ed.). 1991. Handbook of Genetic Algorithms.Van Nostrand Reinhold: New York.

10. Dorsey RE, Johnson JD, Mayer WJ. 1994. A genetic algorithm for training feed forward neural networks. In *Advances in Artificial Intelligence in Economics, Finance and Management* (Vol. 1), Johnson JD, Whinston AB (eds). JAI Press Inc.: Greenwich, CT; 93–111.

11. Dorsey RE, MayerWJ. 1995. Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business and Economic Statistics* 13(1): 53–66.

12. Fahlman SE, Lebiere C. 1990. The cascade-correlation learning architecture. In Advances in Neural Information Processing Systems (Vol. 2), Touretzky DS (ed.). Morgan Kaufmann: San Mateo, CA; 524–532.

13. Funahashi KI. 1989. On the approximate realization of continuous mappings by neural networks. Neural Networks 2(3): 183–192.

14. Guo Zhimao, 2003. An Extensible System for Data Cleaning. Computer Engineer, 29(3), pp. 95-96, 183

15. Heckerling Paul S, Gerber Ben S, 2004. Use of Genetic Algorithms for Neural Networks to Predict Community-Acquired Pneumonia. Artificial Intelligence in Medicine, 30 (1), pp. 71-75.

16. Hornik K, Stinchcombe M, White H. 1989. Multilayer feed-forward networks are universal approximators. Neural Networks 2(5): 359–366.

17. Hsiung JT, SuewatanakulW, Himmelblau DM. 1990. Should backpropagation be replaced by more effective optimization algorithms? Proceedings of the International Joint Conference on Neural Networks (IJCNN) 7: 353–356.

18. Kawabata T. 1991. Generalization effects of k-neighbor interpolation training. Neural Computation 3: 409–417.

19. LeCun Y. 1986. Learning processes in an Asymmetric threshold Network. Disordered Systems and Biological Organizations. Springer-Verlag: Berlin; 233–240.

20. Lenard M, Alam P,Madey G. 1995. The applications of neural networks and a qualitative response model to the auditor’s going concern uncertainty decision. Decision Sciences 26(2): 209–227.

21. Li Mingqiang, 2002. The Principle and Application of Genetic Algorithm. Science Publisher, Beijing.

22. Li Yang, 2004. A Data Mining Architecture Based on ANN and Genetic Algorithm. Computer Engineer, 30(6), pp. 155-156.

23. Madey GR, Denton J. 1988. Credit evaluation with missing fields. *Proceedings of the INNS*, Boston, 456.
24. Masson E, Wang Y. 1990. Introduction to computation and learning in artificial neural networks. *European Journal of Operational Research* 47: 1–28.

25. Michalewicz Z. 1992. Genetic Algorithms + Data Structures = Evolution Programs. Springer: Berlin.

26. Parker D. 1985. Learning logic. Technical report TR-87. Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press: Cambridge MA, 318–362.

27. Prechelt L. 1994. PROBEN1—A set of benchmarks and benchmarking rules for neural network training algorithms. Technical Report 21/94, Fakultat fur Informatik,Universit¨atKarlsruhe,Germany.Anonymous FTP:/pub/papers/techreorts/1994/1994-21.ps.gzon ftp.ira.uka.de.

28. Qing Guofeng, 2003. Acquirement of Knowledge on Data Mining. Computer Engineer, 29(21), pp. 20-22.

29. Rumelhart DE, Hinton GE, Williams RJ. 1986b Learning representations by back propagating errors.. Nature 323: 533–536.

30. Rumelhart DE, Hinton GG, Williams RJ. 1986a. Learning nternal Representations by Error Propagation. Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press: Cambridge MA, 318–362.

31. Schaffer JD, Whitley D, Eshelman LJ. 1992. Combinations of Genetic Algorithms and Neural Networks: A survey of the state of the art, COGANN-92 Combinations of Genetic Algorithms and Neural Networks, *IEEE Computer Society Press*: Los Alamitos, CA; 1–37.

32. Schaffer JD. 1994. Combinations of genetic algorithms with neural networks or fuzzy systems. In Computational Intelligence: Imitating Life, ZuradaJM,

33. Schuster H. 1995. Deterministic Chaos: An Introduction. VCH: Weinheim, New York.

34. Sexton RS, Dorsey RE, Johnson JD. 1998. Toward a global optimum for neural networks: A comparison of the genetic algorithm and backpropagation. Decision Support Systems 22(2): 171–186.

35. Srinivas M., Lalit M.Patnaik, 1994. Genetic Algorithms: *A Survey. IEEE Computer*, 27(6), pp. 17-26.

36. Subramanian V, Hung MS. 1990. A GRG-based system for training neural networks: Design and computational experience. *ORSA Journal on Computing* 5(4): 386–394.

37. Vitthal R, Sunthar P, Durgaprasada Rao Ch. 1995. The generalized proportional-integral-derivative (PID) gradient decent back propagation algorithm. Neural Networks 8(4): 563–569.

38. Wang S. 1995. The unpredictability of standard back propagation neural networks in classification applications. Management Science 41(3): 555–559.

39. Wang Yu, 2005. Predictive Model Based on Improved BP Neural Networks and it's Application. Computer Measurement & Control, 13(1), pp. 39-42.

40. Watrous RL. 1987. Learning algorithms for connections and networks: Applied gradient methods of nonlinear optimization. *Proceedings of the IEEE Conference on Neural Networks* 2, San Diego, 619–627.

41. Werbos P. 1993. The roots of backpropagation: From ordered derivatives to neural networks and political forecasting. JohnWiley: New York.

42. White H. 1987. Some asymptotic results for backpropagation. *Proceedings of the IEEE Conference on Neural Networks* 3, San Diego, 261–266.

43. Xu Lina, 2003. Neural Network Control. *Electronic Industry Publisher*, Beijing.

44. Xu Zezhu, 2004. A Data Mining Algorithm Based on the Rough Sets Theory and BP Neural Network. Computer Engineer and Application, 31, pp. 169-175.

45. Zhang Liming, 1993. The Model and Application of Artificial Neural Network. Fudan University Publisher, Shanghai.

46. A.K.Jain, R.P.W. Duin, and J.Mao, Statistical Pattern Recognition: *A Review, IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22(1), January 2000, pp.4-37.

47. Breitman,L.,Friedman,J.H.,Olshen,R.A.,C.J.,Classification and Regression trees, Wadsworth, Belmont, CA, 1984.

48. Buntine,W.L., Learning classification trees, Statistics and Computing, 1992,pp. 63-73.

49. C. Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

50. Cover, T.M., Hart,P.E., Nearest neighbors pattern classification, *IEEE Trans on Information Theory*, vol. 13, ,1967,pp. 21-27.

51. Hanson R.,Stutz,J.,Cheeseman,P., Bayesian classification with correlation and inheritance, Proceedings of the 12th *International Joint Conference on Artificial Intelligence 2*, Sydney,Australia,Morgan KaufSANN, 1992,pp. 692-698.

52. Michie,D. et al , Machine Learning, Neural and Statistical Classification, Ellis Horwood,1994.

53. R.O.Duda and P.E.Hard, Pattern classification and Scene Analysis, John wiley & Sons, NY, USA, 1973. Richard,M.D, LippSANN,R.P., Neural network classifiers estimate Bayesian a-posterior probabilities, Neural Computation ,vol.3, ,1991,pp. 461-483

54. Tsoi, A.C et al, Comparison of three classification Techniques, CART, C4.5 and multilayer perceptrons , *Advances in Neural Information Processing Systems*, vol. 3, 1991 pp.963-969.

55. V.N.Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of their probabilities, Theory of Probability and its Application, 1971, pp. 264-280.

Authors Profile

Manaswini Pradhan received the B.E. in Computer Science and Engineering, M.Tech in Computer Science from Utkal University, Orissa, India. She is into teaching field from 1998 to till date. Currently she is working as a Lecturer in P.G. Department of Information and Communication Technology, Orissa, India. She is currently persuing the Ph.D. degree in the P.G. Department of Information and communication Technology, Fakir Mohan University, Orissa, India. Her research interest areas are neural networks, soft computing techniques, data mining, bioinformatics and computational biology.

Dr Ranjit Kumar Sahu, M.B.B.S, M.S. (General Surgery), M.Ch. (Plastic Surgery). Worked as an Assistant Surgeon in post doctoral department of Plastic and Reconstructive Surgery, S.C.B. Medical College, Cuttack, Orissa, India. Presently working as a Consultant, Plastic, Cosmetic and Laser Surgery, Mumbai, India, He has five years of research experience in the field of surgery and published many national and international papers in medical field.

Speech Enhancement Using Gradient Based Variable Step Size Adaptive Filtering Techniques

G.V.S.Karthik , M. Ajay Kumar and Md. Zia Ur Rahman

Department of Electronics and Communication Engg.Narasaraopeta Engg. College ,Narasaraopet, 522601, India

E-mail: mdzr_5@ieee.org.

Abstract: *Extraction of high resolution information signals is important in all practical applications. The Least Mean Square (LMS) algorithm is a basic adaptive algorithm has been extensively used in many applications as a consequence of its simplicity and robustness. In practical application of the LMS algorithm, a key parameter is the step size. As is well known, if the step size is large, the convergence rate of the LMS algorithm will be rapid, but the steady-state mean square error (MSE) will increase. On the other hand, if the step size is small, the steady state MSE will be small, but the convergence rate will be slow. Thus, the step size provides a tradeoff between the convergence rate and the steady-state MSE of the LMS algorithm. An intuitive way to improve the performance of the LMS algorithm is to make the step size variable rather than fixed, that is, choose large step size values during the initial convergence of the LMS algorithm, and use small step size values when the system is close to its steady state, which results in Variable Step Size LMS (VSSLMS) algorithms. By utilizing such an approach, both a fast convergence rate and a small steady-state MSE can be obtained. By using this approach various forms of VSSLMS algorithms are implemented. Similar to in the case of the LMS algorithm, a variable step size algorithm is also necessary to obtain both fast convergence rate and small steady state MSE. In this paper various forms of VSSLMS algorithms, which are robust to high variance noise signals are implemented for the construction of adaptive noise cancellers (ANC). Finally we will apply these ANC structures for filtering speech signals. In order to measure the quality of these filters, SNR measurement is considered as quality factor.*

Keywords:*Adaptive filtering, LMS algorithm,MSE,Noise cancellation, Speech enhancement*.

1. Introduction

In real time environment speech signals are corrupted by several forms of noise such as such as competing speakers, background noise, car noise, and also they are subject to distortion caused by communication channels; examples are room reverberation, low-quality microphones, etc. In all such situations extraction of high resolution signals is a key task. In this aspect filtering come in to the picture. Basically filtering techniques are broadly classified as non-adaptive and adaptive filtering techniques. In practical cases the statistical nature of all speech signals is non-stationary; as a result non-adaptive filtering may not be suitable. Speech enhancement improves the signal quality by suppression of noise and reduction of distortion. Speech enhancement has many applications; for example, mobile communications, robust speech recognition, low-quality audio devices, and hearing aids.

Many approaches have been reported in the literature to address speech enhancement. In recent years, adaptive filtering has become one of the effective and popular approaches for the speech enhancement. Adaptive filters permit to detect time varying potentials and to track the dynamic variations of the signals. Besides, they modify their behavior according to the input signal. Therefore, they can detect shape variations in the ensemble and thus they can obtain a better signal estimation. The first adaptive noise cancelling system at Stanford University was designed and built in 1965 by two students. Their work was undertaken as part of a term paper project for a course in adaptive systems given by the Electrical Engineering Department. Since 1965, adaptive noise cancelling has been successfully applied to a number of applications. Several methods have been reported so far in the literature to enhance the performance of speech processing systems; some of the most important ones are: Wiener filtering, LMS filtering [1], spectral subtraction [2]-[3], thresholding [4]-[5]. On the other side, LMS-based adaptive filters have been widely used for speech enhancement [6]–[8]. In a recent study, however, a steady state convergence analysis for the LMS algorithm with deterministic reference inputs showed that the steady-state weight vector is biased, and thus, the adaptive estimate does not approach the Wiener solution. To handle this drawback another strategy was considered for estimating the coefficients of the linear expansion, namely, the block LMS (BLMS) algorithm [9], in which the coefficient vector is updated only once every occurrence based on a block gradient estimation. A major advantage of the block, or the transform domain LMS algorithm is that the input signals are approximately uncorrelated. Recently Jamal Ghasemi et.al [10] proposed a new approach for speech enhancement based on eigenvalue spectral subtraction, in [11] authors describes usefulness of speech coding in voice banking, a new method for voicing detection and pitch estimation. This method is based on the spectral analysis of the speech multi-scale product [12].

In practice, LMS is replaced with its Normalized version, NLMS. In practical applications of LMS filtering, a key parameter is the step size. If the step size is large, the convergence rate of the LMS algorithm will be rapid, but the steady-state mean square error (MSE) will increase. On the other hand, if the step size is small, the steady state MSE will be small, but the convergence rate will be slow. Thus, the step size provides a tradeoff between the convergence rate and the steady-state MSE of the LMS algorithm. The performance of the LMS algorithm may be improved by making the step size variable rather than fixed. The resultant approach with variable step size is known as variable step size LMS (VSSLMS) algorithm [13]. By utilizing such an approach, both a fast convergence rate and a small steady-state MSE can be obtained. Many VSSLMS algorithms are proposed during recent years [14]-[17]. In this paper, we considered the problem of noise cancellation in speech signals by effectively modifying and extending the framework of [1], using VSSLMS algorithms mentioned in [14]-[17]. For that, we carried out simulations on various real time speech signals contaminated with real noise. The simulation results show that the performances of the VSSLMS based algorithms are comparable with LMS counterpart to eliminate the noise from speech signals.

2. Adaptive Algorithms

In this paper we considered various speech signals contaminated with various forms of real noise to demonstrate the concept of adaptive noise cancellation. Figure 1 shows a block schematic of a real transversal FIR filter, here the input values are denoted by u(n), the filter order is denoted by M, and z^{-1} denotes a delay of one sample period. Adaptive filters utilize algorithms to iteratively alter the values of the impulse response vector in order to minimize a value known as the cost function. The cost function, ξ(n), is a function of the difference between a desired output and the actual output of the FIR filter. This difference is known as the estimation error of the adaptive filter, e(n) = d(n) - y(n).

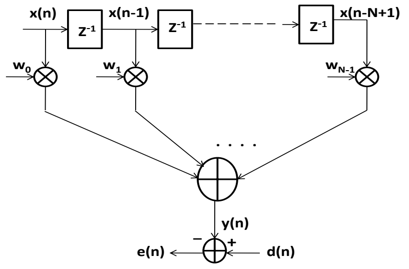


Figure 1: Block diagram of an transversal FIR adaptive filter.

2.1 Basic Adaptive Filtering Structure

Figure 2 shows an adaptive filter with a primary input that is noisy speech signal s_1 with additive noise n_1. While the reference input is noise n_2, which is correlated in some way with n_1. If the filter output is y and the filter error $e = (s_1+n_1)-y$, then

$$\begin{aligned} e^2 &= (s_1+n_1)^2 - 2y\,(s_1+n_1) + y^2 \\ &= (n_1-y)^2 + s_1^2 + 2\,s_1\,n_1 - 2y\,s_1. \qquad (1) \end{aligned}$$

Since the signal and noise are uncorrelated, the mean-squared error (MSE) is

$$E[e^2]=E[(n_1-y)^2]+E[s_1^2] \qquad (2)$$

Minimizing the MSE results in a filter error output that is the best least-squares estimate of the signal s_1. The adaptive filter extracts the signal, or eliminates the noise, by iteratively minimizing the MSE between the primary and the reference inputs. Minimizing the MSE results in a filter error output y that is the best least-squares estimate of the signal s_1.

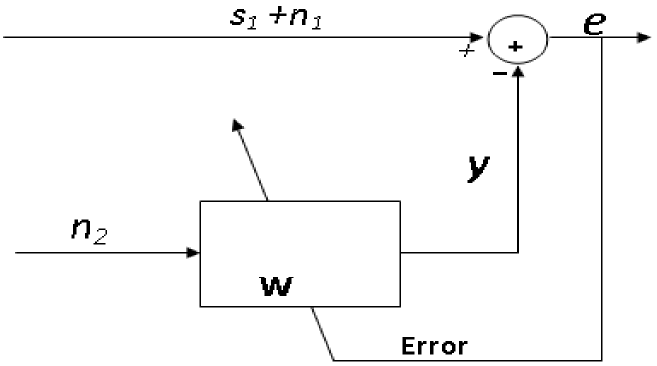


Figure 2: Adaptive Filter Structure.

2.2 Conventional LMS Algorithms

The LMS algorithm is a method to estimate gradient vector with instantaneous value. It changes the filter tap weights so that e(n) is minimized in the mean-square sense. The conventional LMS algorithm is a stochastic implementation of the steepest descent algorithm. It simply replaces the cost function ξ(n) = E[e²(n)] by its instantaneous coarse estimate.

The error estimation e(n) is

$$e(n) = \mathbf{d}(n) - \mathbf{w}(n)\,\mathbf{\Phi}(n) \qquad (4)$$

Coefficient updating equation is

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\,\mathbf{\Phi}(n)\,e(n), \qquad (5)$$

Where μ is an appropriate step size to be chosen as $0 < \mu < \frac{2}{tr\,R}$ for the convergence of the algorithm.

Normalized LMS (NLMS) algorithm is another class of adaptive algorithm used to train the coefficients the adaptive filter.This algorithm takes into account variation in the signal level at the filter output and selecting the normalized step size parameter that results in a stable as well as fast converging algorithm. The weight update relation for NLMS algorithm is as follows

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\,\mathbf{\Phi}(n)\,e(n), \qquad (6)$$

The variable step can be written as,

$$\mu(n) = \mu\,/\,[\,p + \mathbf{\Phi}^{\mathbf{t}}(n)\,\mathbf{\Phi}(n)\,] \qquad (7)$$

Here μ is fixed convergence factor to control maladjustment, μ(*n*) is nonlinear variable of input signal, which changes along with p. The step diminishes and accelerates convergence process. The parameter p is set to avoid denominator being too small and step size parameter too big.

The advantage of the NLMS algorithm is that the step size can be chosen independent of the input signal power and the number of tap weights. Hence the NLMS algorithm has a convergence rate and a steady state error better than LMS algorithm.

2.3 Gradient based variable step size (VSSLMS) LMS Algorithms.

In this paper we considered four types of gradient based VSSLMS algorithms for the implementation of adaptive noise cancellers based on [14]-[17].

2.3.1. Korni's VSSLMS algorithm:

In this algorithm the convergence factor μ is made time-varying in inverse proportion to the input power. As a result this algorithm is shown to be effective for a variety of applications.

$$W_i(n+1) = w_i(n)+\mu(n)e(n)x_i(n) \qquad 0 \leq i \leq N$$

Keep the μ(n) large before the algorithm converges and to reduce it as the algorithm converges. The purpose of the algorithm is to find the minimum of $e^{2,}$ e^2 being a quadratic function of $w_{i,}$ i=0,l,..., N. In other words, the algorithm solves the following simultaneous linear equations:

$$\frac{\partial e^2}{\partial wi} = 0 \qquad 0\leq i \leq N \qquad (8)$$

Since

$$e(n) = d(n)-\sum_{i=0}^{M} W_i(n)x_i(n)$$

Where d(n) is the desired output, eq. (8) can be rewritten in vector form as

$$||eX|| = 0$$

Where ||.|| is the regular vector norm, and

$$X=[x_0, x_1,\ldots\ldots,x_M]$$

μ(n) should be bounded by

$$0 \leq \mu(n) \leq \mu'$$

and if the inputs are identically Gaussian distributed with power σ^2, we have

$$\mu'=1/((M+1)\sigma^2)$$

These discussions suggest a new convergence factor, expressed below:

$$\mu(n)= \mu'(1-e^{-\alpha||e(n)X(n)||}). \qquad (9)$$

Here, α > 0 is the damping parameter. In applications, the norm
||.|| can be replaced by the norm square $||.||^2$

When ||e(n)x(n)|| is large, μ(n)=μ', i.e., the algorithm is in its fast convergence state. After ||e(n)x(n)|| is greatly reduced, μ(n) will be very small, and the algorithm enters its misadjustment minimizing state. Decreasing ||e(n)x(n)|| causes the decreasing of μ(n)**.**Since the misadjustment is directly proportional to μ(n), the misadjustment is thus reduced [18]-[19].

In the case of a non-stationary input, the sudden change of the input induces ||e(n)x(n)|| to become large, which brings the algorithm back to the fast convergence state automatically. It must be pointed out that the "crossover point" of these two states-fast convergence state and misadjustment minimizing state-is governed by the damping parameter α . In fact, there is no clear cut "crossover point," since the exponential function is rather smooth. The larger the parameter α **,** the larger the fast convergence region will be. If α is taken as infinity, then this new algorithm degenerates into the conventional LMS algorithm. As a rule of thumb, α is to be set greater than unity. We note that μ(n) always keeps the algorithm stable.

2.3.2. Kwong's VSSLMS algorithm:

The LMS type adaptive algorithm is a gradient search algorithm which computes a set of weights w_k that seeks to minimize $E(d_k -X^T{}_kW_k)$The algorithm is of the form

$$W_{k+1} = W_k + \mu_k X_k e_k$$

Where

$$e_k = d_k + X^T{}_k W^*{}_k$$

and μ_k is the step size. In the standard LMS algorithm μ_k is a constant. In this μ_k is time varying with its value determined by the number of sign changes of an error surface gradient estimate. Here the new variable step size or VSS algorithm, for adjusting the step size μ_k yields :

$$\mu'_{k+1} = \alpha\mu_k + \gamma e^2_k \qquad 0 < \alpha < 1,\ \gamma > 0$$

and

$$\mu_{k+1} = \begin{cases} \mu_{max} & \text{if } \mu'_{k+1} > \mu_{max} \\ \mu_{min} & \text{if } \mu'_{k+1} < \mu_{min} \\ \mu'_{k+1} & \text{otherwise} \end{cases} \qquad (10)$$

where $0 < \mu_{min} < \mu_{max}$**.** The initial step size μ_0 is usually taken to be μ_{max}, although the algorithm is not sensitive to the choice. The step size μ_k , is always positive and is controlled by the size of the prediction error and the parameters α and γ. Intuitively speaking, a large prediction error increases the step size to provide faster tracking. If the prediction error decreases, the step size will be decreased to reduce the misadjustment. The constant μ_{max} is chosen to ensure that the mean-square error (MSE) of the algorithm remains bounded. A sufficient condition for μ_{max}

$$\mu_{max} \leq 2/(3\ \mathrm{tr}\ (R)) \qquad (11)$$

μ_{min} is chosen to provide a minimum level of tracking ability. Usually, μ_{min} will be near the value of μ that would be chosen for the fixed step size (FSS) algorithm. α must be chosen in the range (0, 1) to provide exponential forgetting.

2.3.3. Mathew's VSSLMS algorithm:

Consider the problem of estimating the desired response signal *d(n)* as a linear combination of the elements of *X(n),* the N-dimensional input vector sequence to the adaptive filter. The popular least mean square (LMS) adaptive filter updates the filter coefficients in the following manner:

$$e(n) = d(n) – X^T(n)H(n)$$

and

$$H(n+1) = H(n) + \mu\ X(n)e(n)$$

Here, $(\cdot)^T$ denotes the matrix transpose of $(\cdot)$, *H(n)* is the coefficient vector at time *n*, and μ is the step-size parameter that controls the speed of convergence as well as the steady-state and/or tracking behavior of the adaptive filter. The selection of μ is very critical for the LMS algorithm. A small μ will ensure small misadjustments in steady state, but the algorithm will converge slowly and may not track the non-stationary behavior of the operating environment very well. On the other hand, a large μ will in general provide faster convergence and better tracking capabilities at the cost of higher misadjustments.

The adaptive step-size algorithm that will be eliminate the "guesswork" involved in selection of the step-size parameter, and at the same time satisfy the following requirements:

1) The speed of convergence should be fast.
2) when operating in stationary environments, the steady-state misadjustment values should be very small. and
3) when operating in non-stationary environments.

The algorithm should be able to sense the rate at which the optimal coefficients are changing and select step-sizes that can result in estimates that are close to the best possible in the mean-squared-error sense. Our approach to achieving the above goals is to adapt the step-size sequence using a gradient descent algorithm so as to reduce the squared-estimation error at each time.

$$e(n) = d(n) – X^T(n)H(n)$$

$$\mu(n) = \mu(n-1) - \frac{\rho}{2}\frac{\partial}{\partial \mu(n-1)}e^2(n)$$

$$=\mu(n-1) - \frac{\rho}{2}\frac{\partial^T e^2(n)}{\partial H(n)} \cdot \frac{\partial H(n)}{\partial \mu(n-1)}$$

$$= \mu(n-1)+\rho e(n)e(n-1)X^T(n-1)X(n) \qquad (12)$$

And

$$H(n+1)= H(n) – \frac{\mu(n)}{2}\frac{\partial e^2(n)}{\partial H(n)}$$

$$= H(n)+\mu(n)e(n)X(n) \qquad (13)$$

In the above equations, ρ is a small positive constant that controls the adaptive behavior of the step-size sequence μ(n).

1) 2.3.4. Aboulnasr's VSSLMS algorithm:

The adaptation step size is adjusted using the energy of the instantaneous error. The weight update recursion is given by

$$W(n+1)= w(n)+\mu(n)e(n)X(n)$$

And updated step-size equation is

$$\mu(n+1)=\alpha\mu(n)+\gamma e^2(n) \qquad (14)$$

where $0<\alpha<1, \gamma>0$, and μ(n+1) is set to or when it falls below or above these lower and upper bounds, respectively. The constant μ_{max} is normally selected near the point of instability of the conventional LMS to provide the maximum possible convergence speed. The value of μ_{max} is chosen as a compromise between the desired level of steady state misadjustment and the required tracking capabilities of the algorithm. The parameter γ controls the convergence time as well as the level of misadjustment of the algorithm. At early stages of adaptation, the error is large, causing the step size to increase, thus providing faster convergence speed. When the error decreases, the step size decreases, thus yielding smaller misadjustment near the optimum. However, using the instantaneous error *energy* as a measure to sense the state of the adaptation process does not perform as well as expected in the presence of measurement noise. The output error of the identification system is

$$e(n)=d(n)-X^T(n)W(n)$$

where d(n) is the desired signal is given by

$$d(n)=X^T(n)W^*(n)+\xi(n) \qquad (15)$$

ξ(n) is a zero-mean independent disturbance, and $W^*(n)$ is the time-varying optimal weight vector. Substituting (3) and (4) in the step-size recursion, we get

$$\mu(n+1)=\alpha\mu(n)+\gamma\ V^T(n)X(n)X^T(n)V(n) +\gamma\xi^2(n)-2\gamma\xi(n)V^T(n)X(n) \qquad (16)$$

Where $V(n)=W(n)-W^*(n)$ is the weight error vector. The input signal autocorrelation matrix, which is defined as $R=E\{X(n)X^T(n)\}$, can be expressed as $R=Q\Lambda Q^T$, where Λ is the matrix of eigenvalues, and Q is the model matrix of

R. using $V'(n)=Q^TV(n)$ and $X'(n) = Q^TX(n)$, then the statistical behavior of $\mu(n+1)$ is determined.

$$E\{\mu(n+1)\}=\alpha E\{\mu(n)\}+\gamma(E\{\xi^2(n)\}+E\{V'^T(n)\ \Lambda V'(n)\})$$

where we have made use of the common independence assumption of $V'(n)$ and $X'(n)$. Clearly, the term $E\{V'^T(n)\ \Lambda V'(n)\}$ influences the proximity of the adaptive system to the optimal solution, and $\mu(n+1)$ is adjusted accordingly. However, due to the presence of $E\{\xi^2(n)\}$, the step-size update is not an accurate reflection of the state of adaptation before or after convergence. This reduces the efficiency of the algorithm significantly. More specifically, close to the optimum, $\mu(n)$ will still be large due to the presence of the noise term $E\{\xi^2(n)\}$. This results in large misadjustment due to the large fluctuations around the optimum. In this paper, a different approach is proposed to control step-size adaptation. The objective is to ensure large $\mu(n)$ when the algorithm is far from the optimum with $\mu(n)$ decreasing as we approach the optimum *even in the presence of this noise*. The proposed algorithm achieves this objective by using an estimate of the autocorrelation between e(n) and e(n-1) to control step-size update. The estimate is a time average e(n)e(n-1) of that is described as

$$p(n)=\beta p(n-1)+(1-\beta)e(n)e(n-1)$$

The use of p(n) in the update of $\mu(n)$ serves two objectives. First, the error autocorrelation is generally a good measure of the proximity to the optimum. Second, it rejects the effect of the uncorrelated noise sequence on the step-size update. In the early stages of adaptation, the error autocorrelation estimate $p^2(n)$ is large, resulting in a large $\mu(n)$. As we approach the optimum, the error autocorrelation approaches zero, resulting in a smaller step size. This provides the fast convergence due to large initial $\mu(n)$ while ensuring low misadjustment near optimum due to the small final $\mu(n)$ even in the presence of $\xi(n)$. Thus, the proposed step size update is given by

$$M(n+1)= \alpha\mu(n)+\gamma p(n)^2$$

The positive constant $\beta(0<\beta<1)$ is an exponential weighting parameter that governs the averaging time constant, i.e., the quality of the estimation. In stationary environments, previous samples contain information that is relevant to determining an accurate measure of adaptation state, i.e.,the proximity of the adaptive filter coefficients to the optimal ones. Therefore, β should be $\approx$1. For non stationary optimal coefficients, the time averaging window should be small enough to allow for forgetting of the deep past and adapting to the current statistics, i.e., $\beta<1$. The step size can be rewritten as

$$\mu(n+1)=\alpha\mu(n)+\gamma[E\{V^T(n)X(n)X^T(n-1)V(n-1)\}]^2. \quad (17)$$

It is also clear from above discussion that the update of $\mu(n)$ is dependent on how far we are from the optimum and is not affected by independent disturbance noise. Finally, the considered algorithm involves two additional update equations compared with the standard LMS algorithm. Therefore, the added complexity is six multiplications per iteration. These multiplications can be reduced to shifts if the parameters α,β,γ, are chosen as powers of 2. A summary of step size update equation is shown in Table I.

Table I: Summary of all VSSLMS algorithms.

Name of the algorithm	Update of the step size
Karin's VSSLMS	$\mu(n)= \mu^1(1-e^{-\alpha\|e(n)X(n)\|})$ $\mu^1=1/((M+1)\sigma^2)$
Kwong's VSSLMS	$\mu^{\mathbf{1}}_{k+1} = \alpha\mu_k + \gamma e^2_k$
Mathew's VSSLMS	$\mu(n) = \mu(n-1)+\rho e(n)e(n-1)X^T(n-1)X(n)$
Aboulnasr's VSSLMS	$\mu(n+1)=\alpha\mu(n)+\gamma[E\{V^T(n)X(n)X^T(n-1)V(n-1)\}]^2$

The performance of these algorithms compared from the convergence characteristics shown in figure 3. From the convergence curves it is clear that the performance of VSSLMS algorithms is better than the conventional LMS / NLMS algorithms. Among the four VSSLMS algorithms Aboulnsr's algorithm is better than the other. From the figure it is clear that the VSSLMS algorithms converge very slowly at the beginning, but speed up as the MSE level drops.

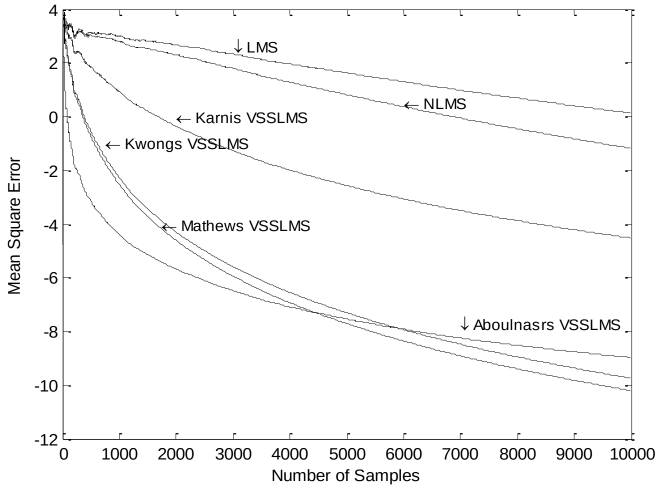


Figure 3: Convergence Characteristics of various algorithms.

3. Simulation Results

To show that VSS LMS algorithms are appropriate for speech enhancement we have used real speech signals and real noisy signals. These real speech signals are shown in figure 4. The sample-I is a practically recorded signal with 53569 samples. Sample-II is obtained from database and it has 68689 samples. Sample-III has 48136 samples, sample-IV is a real signal with 50000. These are shown in figure 4. In the figure *number of samples* is taken on *x-axis* and *amplitude* is taken on *y-axis*.

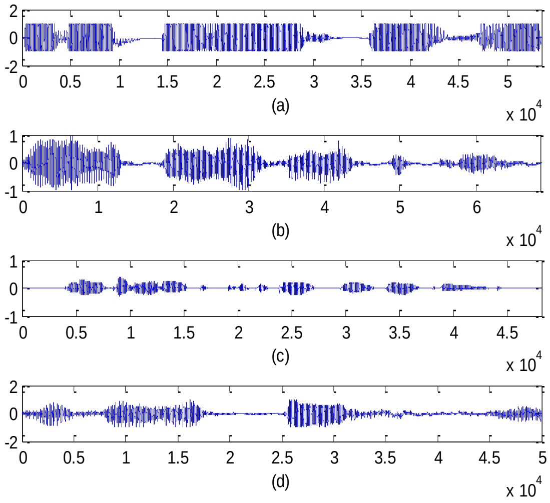


Figure 4: Real Speech Signals (a). Sample-I, (b). Sample-II,(c). Sample-III,(d). sample-IV.

To evaluate the performance of the adaptive algorithms and to prove the non-stationary tracking performance of the algorithms, both synthetic and real noises are taken. Some noises are shown in figure 5.

3.1 Characteristics of FIR filter

For the implementation of adaptive noise canceller we have chosen a second order FIR filter. The considered filter is a direct form II stable filter. The numerator length is two, denominator length is three, number of multipliers are two, number of adders is one, number of states are two, multiplications per input sample are two, additions per input sample is one. The transfer function of the filter is given by,

$$H(z) = 2Z^2 – 5Z+2 \ / 2Z^2(Z-1).$$

The magnitude – phase response, pole-zero plot and impulse response of the considered FIR filter are shown in figure 6(a), 6(b) and 6(c) respectively.

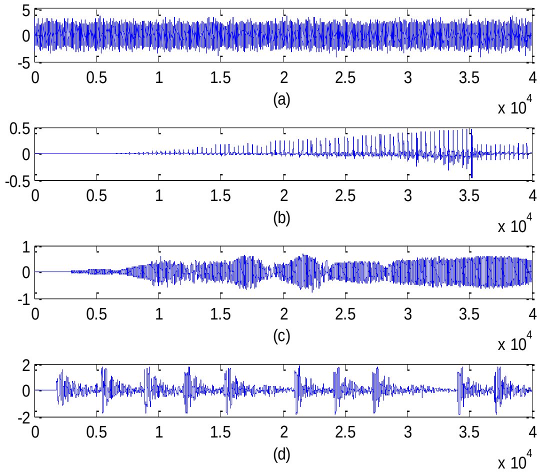


Figure 5: Synthetic and real noises used in this paper (a). Random noise (b). High voltage spark noise, (c). Speaker noise, (d). Battle field noise.

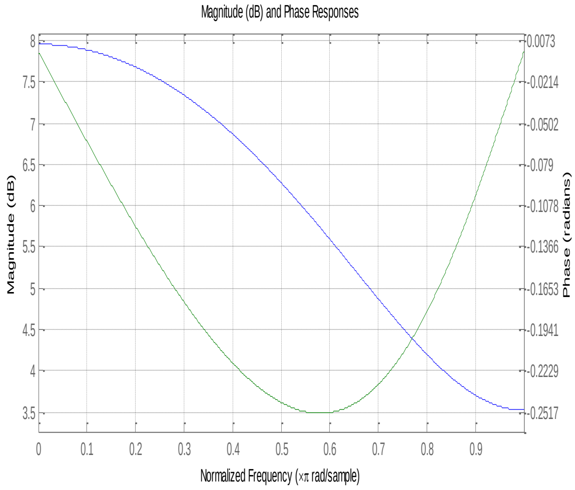


Figure 6(a): Magnitude and Phase response of the FIR filter.

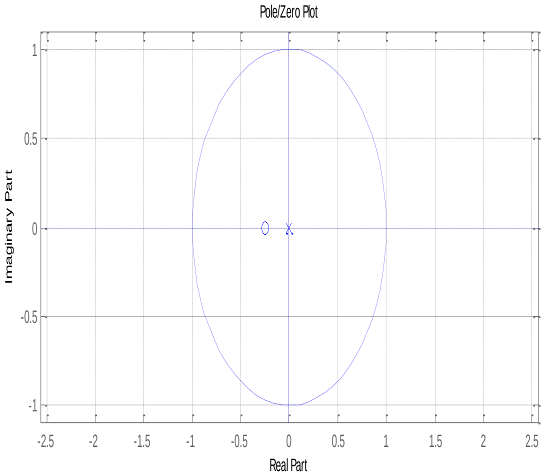


Figure 6(b): Pole Zero plot of the FIR filter.

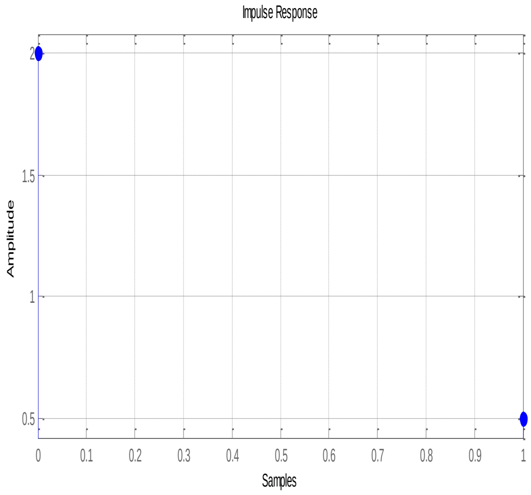


Figure 6(c): Impulse response of the FIR filter.

3.2. Simulation Results for Random Noise removal

As a first step in adaptive noise cancellation application, the speech signal corresponding to sample-I is corrupted with random noise and is given as input signal to the adaptive filter shown in figure 2. As the reference signal must be somewhat correlated with noise in the input, the random noise signal is given as reference signal. The filtering results are shown in figures 7 and 8. To evaluate the performance of the algorithms signal-to-noise (SNR) improvement is measured and tabulated in Table II.

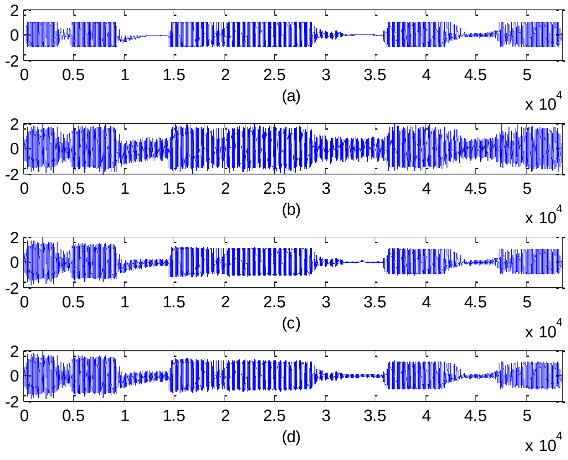


Figure 7**:**Typical filtering results of random noise removal (a) Original Speech Signal, (b) noisy signal, (c) recovered signal using LMS algorithm, (d) recovered signal using NLMS algorithm.

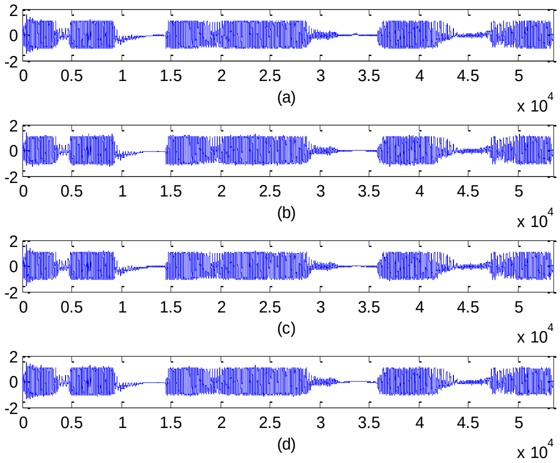


Figure 8**:** Typical filtering results of random noise removal (a) recovered signal using Karni's VSSLMS algorithm, (b) recovered signal using Kwongi's VSSLMS algorithm, (c) recovered signal using Mathew's VSSLMS algorithm, (d) recovered signal using Aboulnasr's VSSLMS algorithm.

3.3. Adaptive cancellation of real high voltage murmuring

In this experiment a speech signal corresponding to sample-II contaminated with high voltage murmuring is given as in put to the filter. The filtering results are shown in figures 9 and 10. The SNR contrast is shown in Table-II.

3.4. Simulation Results for battle field noise removal

In this experiment the speech signal contaminated with a real battle field noise (gun firing noise predominates in this noise) is given as input to the adaptive filter shown in figure 2. As the reference signal must be somewhat correlated with noise in the input, the noise signal is given as reference signal. The filtering results are shown in figures 11 and 12. To evaluate the performance of the algorithms signal-to-noise (SNR) improvement is measured and tabulated in Table II.

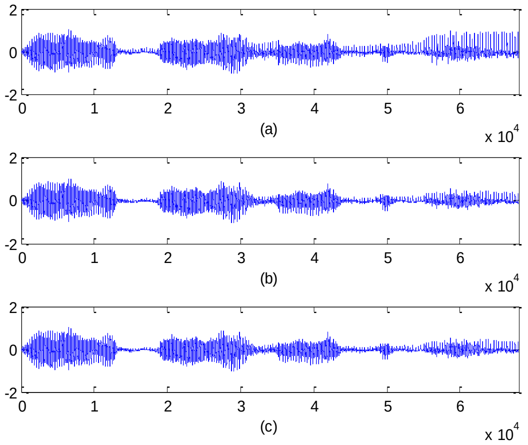


Figure 9**:** Typical filtering results of high voltage noiseremoval (a) Speech signal with high voltage noise,

(b) recovered signal using LMS algorithm, (c) recovered signal using NLMS algorithm.

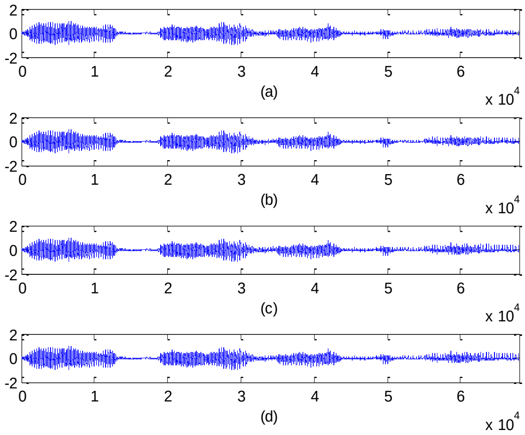


Figure 10: Typical filtering results of high voltage noise removal (a) recovered signal using Karni's VSSLMS algorithm, (b) recovered signal using Kwongi's VSSLMS algorithm, (c) recovered signal using Mathew's VSSLMS algorithm, (d) recovered signal using Aboulnasr's VSSLMS algorithm.

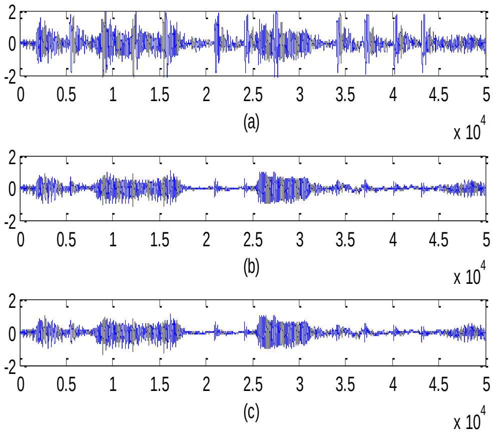


Figure 11: Typical filtering results of battle field noise removal (a) Speech signal with battle field noise, (b) recovered signal using LMS algorithm, (c) recovered signal using NLMS algorithm.

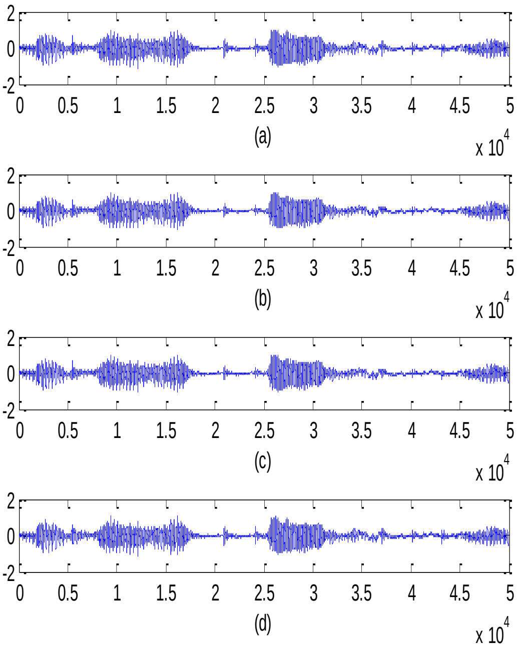


Figure 12: Typical filtering results of battle field noise removal (a) recovered signal using Karni's VSSLMS algorithm, (b) recovered signal using Kwongi's VSSLMS algorithm, (c) recovered signal using Mathew's VSSLMS algorithm, (d) recovered signal using Aboulnasr's VSSLMS algorithm.

3.5. Simulation Results for speaker noise removal

In this case speech signal contaminated with a loud speaker is given as input to the adaptive filter shown in figure 2. As the reference signal must be somewhat correlated with noise in the input, the noise signal is given as reference signal. The filtering results are shown in figures 13 and 14. To evaluate the performance of the algorithms signal-to-noise (SNR) improvement is measured and tabulated in Table II.

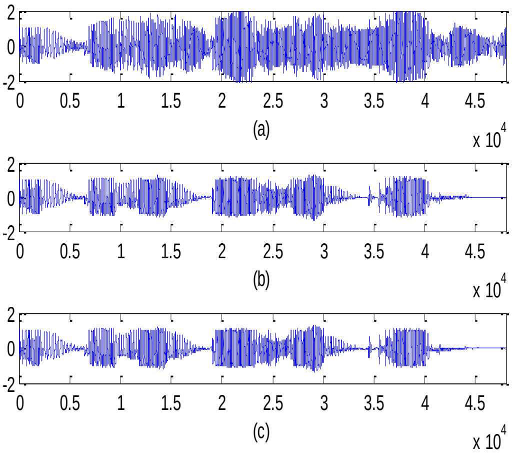


Figure 13: Typical filtering results of speaker noise removal (a) Speech signal with speaker noise, (b) recovered signal using LMS algorithm, (c) recovered signal using NLMS algorithm.

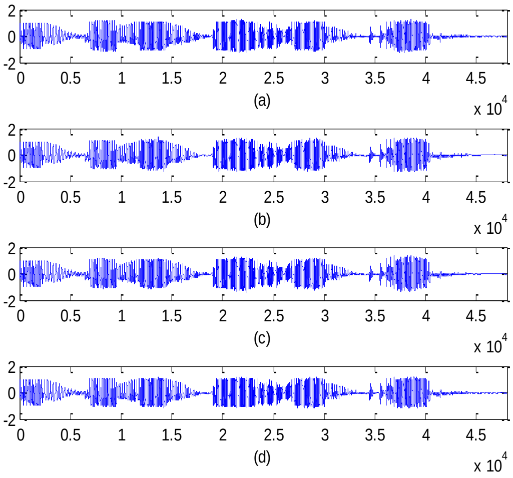


Figure 14**:** Typical filtering results of speaker noise removal (a) recovered signal using Karni's VSSLMS algorithm, (b) recovered signal using Kwongi's VSSLMS algorithm, (c) recovered signal using Mathew's VSSLMS algorithm, (d) recovered signal using Aboulnasr's VSSLMS algorithm.

Table II: SNR contrast of various algorithms.

Sample Number	SNR Imp. after LMS Filtering	SNR Imp. after NLMS Filtering	SNR Imp. after Karni's VSSLMS Filtering	SNR Imp. after Kwongs's VSS LMS Filtering	SNR Imp. after Mathew's VSS LMS Filtering	SNR Imp.after Aoulnasr's VSSLMS Filtering
Sample I	7.9095	6.3031	11.7786	11.9863	13.8720	15.7111
Sample II	8.2398	7.8904	11.9872	12.2897	13.5021	15.3921
Sample III	6.4891	6.6937	12.0739	12.9541	13.7289	14.9231
Sample IV	7.3642	7.7183	12.4923	13.1054	14.0625	15.8232

4. Conclusion

In this paper the problem of noise removal from speech signals using VSSLMS based adaptive filtering is presented. For this, the same formats for representing the data as well as the filter coefficients as used for the LMS algorithm were chosen. As a result, the steps related to the filtering remains unchanged. The proposed treatment, however exploits the modifications in the weight update formula for all categories to its advantage and thus pushes up the speed over the respective LMS-based realizations. Our simulations, however, confirm that the ability of VSSLMS algorithms is better than conventional LMS and NLMS algorithms in terms of SNR improvement and convergence rate. Hence these algorithms are acceptable for all practical purposes.

5. REFERENCES

[1] B. Widrow, J. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H.Hearn, J. R. Zeidler, E. Dong, and R. Goodlin,"Adaptive noise cancelling: Principles and applications ", Proc. IEEE, vol. 63, pp.1692-1716, Dec. 1975.

[2] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," IEEE Trans. On Speech and Audio Processing, vol. 6, pp. 328-337, 1998.

[3] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," Speech Communication, vol. 24, pp. 249-257, 1998.

[4] H. Sheikhzadeh, and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," Proc. of the Eurospeech, 2001.

[5] S. Salahuddin, S. Z. Al Islam, M. K. Hasan, M. R. Khan, "Soft thresholding for DCT speech enhancement," Electron. Letters, vol. 38,
no.24, pp. 1605-1607, 2002.

[6] J. Homer, "Quantifying the convergence speed of LMS adaptive filter with autoregressive inputs," Electronics Letters, vol. 36, no. 6, pp. 585– 586, March 2000.

[7] H. C. Y. Gu, K. Tang and W. Du, "Modifier formula on mean square convergence of LMS algorithm," Electronics Letters, vol. 38, no. 19, pp. 1147 –1148, Sep 2002.

[8] M. Chakraborty and H. Sakai, "Convergence analysis of a complex LMS algorithm with tonal reference signals," IEEE Trans. on Speech and Audio Processing, vol. 13, no. 2, pp. 286 – 292, March 2005.

[9] S. Olmos , L. Sornmo and P. Laguna, ``Block adaptive filter with deterministic reference inputs for event-related signals:BLMS and BRLS," IEEE Trans. Signal Processing, vol. 50, pp. 1102-1112, May.2002.

[10] Jamal Ghasemi and Mohammad Reza Karami Mollaei, "A New Approach for Speech Enhancement Based On Eigenvalue Spectral Subtraction", Signal Processing: An International Journal, vol. 3, Issue. 4, pp. 34-41.

[11] Mohamed Anouar Ben Messaoud, Aïcha Bouzid and Noureddine Ellouze," A New Method for Pitch Tracking and Voicing Decision Based on Spectral Multi-scale Analysis", Signal Processing: An International Journal, vol. 3, Issue. 5, pp. 144-152

[12] M.Satya Sai Ram, P. Siddaiah and M. Madhavi Latha," USEFULNESS OF SPEECH CODING IN VOICE BANKING", Signal Processing: An International Journal, vol. 3, Issue. 4, pp. 42-52.

[13] Yonggang Zhang, Ning Li, Jonathon A. Chambers, and Yanling Hao, "New Gradient-Based Variable Step Size LMS Algorithms," EURASIP Journal on Advances in Signal Processing vol. 2008, Article ID 529480, 9 pages, doi:10.1155/2008/529480.

[14] S. Karni and G. Zeng, "A new convergence factor for adaptive filters," *IEEE Transactions on Circuits and Systems*, vol. 36, no. 7, pp. 1011–1012, 1989.

[15] R. H. Kwong and E.W. Johnson, "A variable step-size LMS algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 7, pp. 1633–1642, 1992.

[16] V. J. Mathews and Z. Xie, "A stochastic gradient adaptive filter with gradient adaptive step-size," *IEEE Transactions on Signal Processing*, vol. 41, no. 6, pp. 2075–2087, 1993.

[17] T. Aboulnasr and K.Mayyas, "A robust variable step-size LMStype algorithm: analysis and simulations," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 631–639, 1997.

[18] B. Widrow and S. D. Stearns, *Aduptiue Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985

[19] G. Zeng, "A new adaptive IIR algorithm and the convergence factors for digital and analog adaptive filters," Ph.D. dissertation, University of NewMexico, May 1988.

[20] Farhang-Boroujeny, B., "Adaptive Filters- Theory and applications", John Wiley and Sons, Chichester, UK, 1998.

High Performance Human Face Recognition using Independent High Intensity Gabor Wavelet Responses: A Statistical Approach

Arindam Kar[1], Debotosh Bhattacharjee[2], Dipak Kumar Basu[2*], Mita Nasipuri[2], Mahantapas Kundu[2]

[1] Indian Statistical Institute Kolkata-700108, India,
[2]Department of Computer Science and Engineering, Jadavpur University, Kolkata- 700032, India,
* AICTE Emeritus Fellow.
Email: {kgparindamkar@gmail.com, debotosh@indiatimes.com, dipakkbasu@gmail.com, mita_nasipuri@gmail.com, mkundu@cse.jdvu.ac.in}

Abstract: *In this paper, we present a technique by which high-intensity feature vectors extracted from the Gabor wavelet transformation of frontal face images, is combined together with Independent Component Analysis (ICA) for enhanced face recognition. Firstly, the high-intensity feature vectors are automatically extracted using the local characteristics of each individual face from the Gabor transformed images. Then ICA is applied on these locally extracted high-intensity feature vectors of the facial images to obtain the independent high intensity feature (IHIF) vectors. These IHIF forms the basis of the work. Finally, the image classification is done using these IHIF vectors, which are considered as representatives of the images. The importance behind implementing ICA along with the high-intensity features of Gabor wavelet transformation is twofold. On the one hand, selecting peaks of the Gabor transformed face images exhibit strong characteristics of spatial locality, scale, and orientation selectivity. Thus these images produce salient local features that are most suitable for face recognition. On the other hand, as the ICA employs locally salient features from the high informative facial parts, it reduces redundancy and represents independent features explicitly. These independent features are most useful for subsequent facial discrimination and associative recall. The efficiency of IHIF method is demonstrated by the experiment on frontal facial images dataset, selected from the FERET, FRAV2D, and the ORL database.*

Keywords: *Feature extraction; Gabor Wavelets; independent high-intensity feature (IHIF); Independent Component Analysis (ICA); Specificity; Sensitivity; Cosine Similarity Measure.*

1. Introduction

Face authentication has gained considerable attention recently, through the increasing need for access verification systems. Such systems are used for the verification of a user's identity on the Internet, when using a bank automaton, when entering to secured building, etc. Face recognition involves computer recognition of personal identity based on geometric or statistical features derived from face images. Even though humans can identify faces with ease, but building such an automated system that accomplishes such objectives is, very challenging. The challenges are even more intense when there are large variations due to illumination conditions, viewing directions or poses, facial expression, aging, etc. The Face recognition research provides the cutting edge technologies in airports, government, military facilities, countries borders, and so on. Principal component analysis (PCA) is a popular statistical method to find useful image representations. Independent Component Analysis (ICA) has emerged over the years as one powerful solution to the problem of blind source separation [1], [2], [3], [4], [5], [6]. Turk and Pentland [7] developed a well-known Eigenfaces method, which sparked great interest in applying statistical techniques to face recognition. While PCA considers the second-order moments only and it un-correlates the data, but ICA would further reduce statistical dependencies and produce a sparse and independent code useful for subsequent pattern discrimination and associative recall [8]. The metric induced by ICA is superior to PCA in the sense that it may provide a representation more robust to the effect of noise [9]. It is, therefore, possible for ICA to be better than PCA for reconstruction in noisy or limited precision environments. When the sources models are sparse, ICA is closely related to the so called non-orthogonal "rotation" methods in PCA and factor analysis. The goal of these rotation methods is to find directions with high concentrations of data. ICA can be used to find interesting non-orthogonal "rotations" [10, 11, 12]. One of the most successful face recognition methods is based on graph matching of coefficients proposed by Lades et. al. [13], which are obtained from Gabor wavelet (GW) responses. However, such graph matching algorithm methods have some disadvantages due to their matching complexity, manual localization of training graphs overall execution time, and extraction of special characteristics of each individual. Bell and Sejnowski [14] has shown that image bases that produce independent outputs from natural scenes are local oriented spatially filters similar to the response properties of simple cells. Conversely, it has been seen that Gabor filters, closely model the responses of simple cells, separate higher-order dependencies [14, 15, 16]. We have already done a work [17], the algorithm proposed for the extraction of high-energized points from Gabor wavelet transforms, has the time complexity of the order $O(N_1 \cdot N_2 \cdot n)$ where N1, N2 is the width and height of the image and n is the number of GW response. If N1=N2= n, then the time complexity is of the order $O(n^3)$. The proposed algorithm shown in section III has total computation time $[(5N_1+1)+(5N_1 \cdot (8N_2+1))+5N_1 \cdot 8N_2]$, so the time complexity is $O(N_1 \cdot N_2)$. Thus if N1=N2= n, the time complexity is of the order $O(n^2)$. The method introduced in this paper differs from the one in [18], as the latter method integrated the independent properties of only the high-energized feature vectors, which is a nontrivial challenge for implementing fast and automated face recognition systems. As PCA is only sensitive to the power spectrum of images, so it might not be well suited for representing natural images. In particular, it has been observed that images are better described as linear combinations of sources with long tailed distributions [19]. Recent approaches of image representation emphasize on data-driven learning-based techniques, such as the statistical modeling methods [9, 10, 11] the neural network-based learning methods [12], the statistical learning theory and Support vector machine (SVM) based methods [20, 21]. The motivation of using feature based methods is due to their representation of the face image in a very compact way and hence lowering the memory needs.

In this paper, a robust and reliable automatic IHIF method for face recognition is proposed, which is robust to occlusion and illumination changes, and can overcome those disadvantages. This solution is based on selecting intensity peaks of GW responses for the facial high–intensity feature vector construction, instead of using predefined graph nodes as in elastic graph matching (EGM) [13], which reduces representative capability of Gabor wavelets.

The rest of this paper is organized as follows: Section 2 describes Gabor wavelet convolution outputs of facial images. Section 3 describes extraction of high-intensity feature vector from the convolution outputs of Gabor transformed images. Section 4 the application of ICA on the extracted high-intensity feature vector, Section 5 assesses the performance of the IHIF method on the face recognition task by applying it on FERET[22], FRAV2D [23], and the ORL [24] database and comparing with some popular face recognition schemes, and finally we conclude in Section 6.

2. 2D Gabor Wavelet Analysis

Physiological studies found simple cells, in human visual cortex, that are selectively tuned to orientation, as well as to spatial frequency, could be approximated by 2D Gabor filters [25, 26, 27]. It has been shown that using Gabor filters as front-end of an automated face recognition system are highly successful [28, 29]. 2D Gabor functions are similar to enhancing edge contours, as well as valleys and ridge contours of the image. The Gabor wavelets, exhibit strong characteristics of spatial locality and orientation selectivity, and are optimally localized in the spatial and frequency domain [26]. The Gabor filter used here has the following general form:

$$\varphi_{\mu,\nu}(z)=\frac{\left|k_{\mu,\nu}\right|^2}{\sigma^2}e^{-\left\|k_{\mu,\nu}\right\|^2\left\|z\right\|^2/2\sigma^2}\left[e^{ik_{\mu,\nu}z}-e^{-\left(\sigma^2/2\right)}\right] \quad (1)$$

The center frequency of i^{th} filter is given by the characteristic wave vector

$$\vec{k}_i=\begin{bmatrix}\mathrm{k}\ jx\\ \mathrm{k\ jy}\end{bmatrix}=\begin{bmatrix}\mathrm{k}_\upsilon\ \cos\theta_\mu\\ k_\upsilon\ \sin\theta_\mu\end{bmatrix},$$

having a scale and orientation given by $(\mathrm{k}_\mathrm{v},\theta_\mu)$, where μ and ν defines the orientation and scale of Gabor kernels, $z=(x,y)$, is the variable in spatial domain, $\|\cdot\|$ denotes the norm, and $\mathrm{k}_{\mu,\nu}$ is the frequency vector which determines the scale and orientation of Gabor kernels, where $k_\nu = k_{\max}/f^\nu$ and $\mathrm{k}_{\max}=\pi/2$, $\phi_\mu=\pi\mu/8$, $\mu=0,..7$, where f is the spacing factor between kernels in the frequency domain. Convolving the image with complex Gabor filters with 5 spatial frequency, and eight orientations, captures the whole frequency spectrum, both amplitude and phase $\mathrm{O}_{\mu,\upsilon}(\mathrm{z})$is denoted as the magnitude of the convolution outputs. The term $\mathrm{e}^{-\left(\sigma^2/2\right)}$ is subtracted from equation (1) in order to make the kernel DC-free, thus become insensitive to illumination. The decomposition of an image *I* into these states is called wavelet transformation of the image:

$$R_i(\bar{x})=\int I(\bar{x}')\varphi_i(\bar{x}-\bar{x}')d\bar{x}' \quad (2)$$

where $\mathrm{I}(\overline{\mathrm{x}}')$ is the image intensity value at $\bar{x}$, $\mathrm{i}=1,\dots,40$. Fig. 1(a) shows the real part of Gabor kernels , and Fig. 2(b) shows their magnitude with the following parameters : $\sigma=2\pi$, $\mathrm{k}_{\max}=\pi/2$, and $f=\sqrt{2}$, at 5 different scales and 8 orientations respectively.

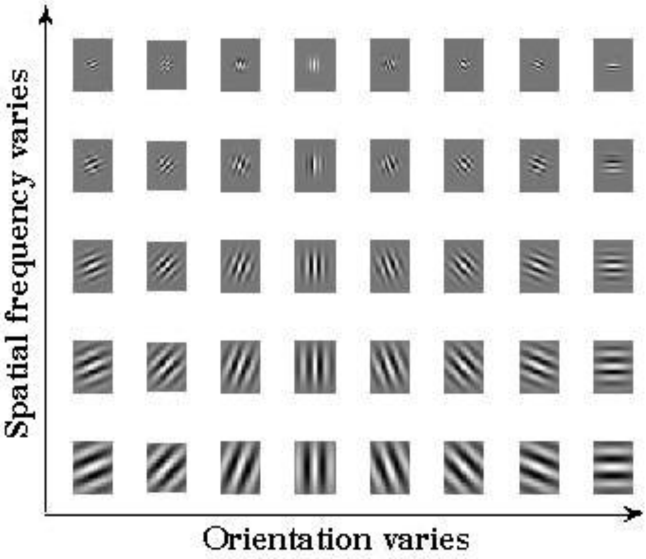


Fig. 1. (a) Real part of the Gabor kernels

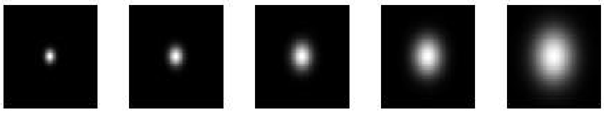


Fig.1 (b) Magnitudes of the Gabor kernels at 5 different frequency and 8 orientations

3. High-Intensity feature extraction

The Gabor wavelet representation of an image is the convolution of the image with a family of Gabor kernels as defined by (1). Let $\mathrm{I}(\mathrm{x},\mathrm{y})$ be the gray level distribution of an image, the convolution output of image I and a Gabor kernel $\varphi_{\mu,\upsilon}$ is defined as follows:

$$O_{\mu,\upsilon}(z)=\mathrm{I}(z)*\varphi_{\eta,\upsilon}(z) \quad (3)$$

where $z=(\mathrm{x},\mathrm{y})$, and * denotes the convolution operator. Applying the convolution theorem, we can derive the convolution output from (3) via the fast Fourier transform (FFT)

$$\Im\{O_{\mu,\upsilon}(z)\}=\Im\{I(z)\}\Im\{\varphi_{\mu,\upsilon}(z)\} \quad (4)$$

$$O_{\mu,\upsilon}(z)=\Im^{-1}\{\Im\{I(z)\}\Im\{\varphi_{\mu,\upsilon}(z)\}\} \quad (5)$$

where $\Im$ and $\Im^{-1}$ denote the Fourier and inverse Fourier transform, respectively. The convolution outputs (the magnitude) of a sample image and the Gabor kernels (see fig. 1. (a)) is shown in Fig. 2.

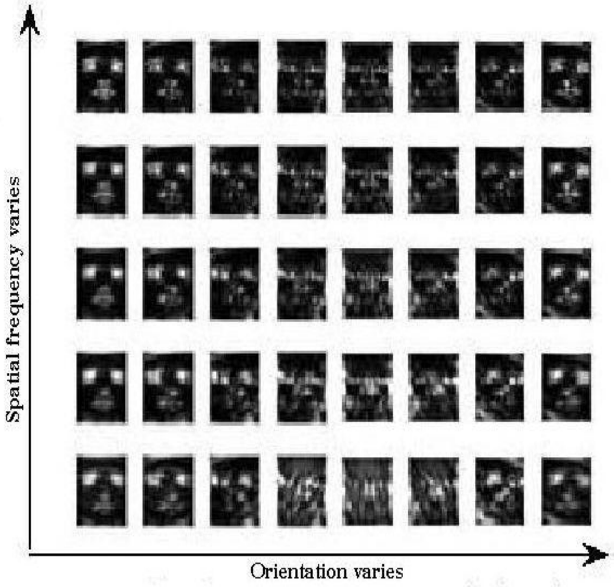


Fig.2 High intensity Gabor filter responses of a face image

The outputs exhibits strong characteristics of special locality, scale and orientation selectivity, such characteristics produce salient local features such as the eyes, nose, mouth, scars and dimples that are suitable for visual event recognition, and hence making them a suitable choice for feature extraction of images. As the outputs $O_{\mu,\upsilon}(z)$ consist of different local, scale, and orientation features, all these features are concatenated into a single column vector in order to derive a feature vector V. The feature extraction of the proposed method has two main steps: (a) High intensity feature point localization (b) Feature vector generation. It is to be noted that we applied the magnitude but did not use the phase, which is considered to be consistent with the application of Gabor representations in [15, 16]. As the outputs $(O_{\mu,\nu}(z):\mu\in\{0,...,4\},\nu\in\{0,...7\})$ consists of 40, different local, scale and orientation features, the dimensionality of the Gabor transformed image space is very high. So the following technique is applied for the extraction of low dimensional high intensity vector X_k from the convolution outputs. (a) The algorithm for high intensity feature point localization is as follows:

Start

Step 1: Find G_{ij} = GWT of the image I of size (M×N) where i=0,…,4;j=0,…,7. are the scale and orientation variation.

Step 2: Find I_{ij} = modulus of G_{ij}.

Step 3: Find M_{ij} =mean of I_{ij}.

Step 4: Divide the matrix I_{ij} into square blocks of size (W×W).Thus total no. of blocks $=\left\lfloor\frac{M}{W}\right\rfloor\times\left\lfloor\frac{N}{W}\right\rfloor$, denote it by B_{ijk} where k=1,…, $\left\lfloor\frac{M}{W}\right\rfloor\times\left\lfloor\frac{N}{W}\right\rfloor$.

Step 5: Find vector χ_{ijk} such that it contains pixel values of B_{ijk}, which are greater than M_{ij}. If $\chi_{ijk}=[\ ]$, then put $\chi_{ijk}=M_{ij}$.

Step 6: For t=1 to length of (χ_{ijk})

if $\left|\chi_{ijkt}-\overline{\chi}_{ijk}\right|<\mathit{threshold}$,

take $\mathrm{Y}_{\mathrm{ijk}}=[\,\chi_{\mathrm{ijkt}}\,]$,where $\overline{\chi}_{ijk}$ =mean of χ_{ijk}.

if $Y_{ijk}=[\ ]$, then $Y_{ijk}=\overline{\chi}_{ijk}$.Threshold is taken as 3.

Step 7: Z_{ijk}=Sort (Y_{ijk}) descending order of the values.

Step 8: L= $\underset{k}{Minimum}(length\ of\,(Z_{ijk}))$

Step 9: Select the top L elements of every vector Z_{ijk} and store it in a column vector $\chi_{I_{ij}}$ .Thus we get high intensity feature vector $\chi_{I_{ij}}$.

Step 10: Repeat step 2 to step 9 for all I_{ij} of image I.

end

(b) Feature vector generation: The final high intensity feature vector χ_k is generated by accumulation of the elements of $(\chi_{I_{ij}})$ column wise to a single vector for the kth individual faces. The advantage of this feature extraction over, the previous algorithm [17] is that, previously the computation was dependent on the number of the GW responses, which is eliminated here. Thus the time complexity of this new algorithm is reduced by a factor of 'λ' where λ is the number of GW responses.

4. ICA on the extracted High-Intensity Feature Vectors of Gabor responses

As ICA would further reduce redundancy and represent independent features explicitly [20]. So independence property of these high-intensity features facilitates the application for image classification. These images, thus, produce salient local features that are most suitable for face recognition. On the other hand, we have input representations that are robust to noise and better reflect the data.

4.1 Preprocessing for ICA

Features of facial images are obtained through eight directions and five scales respectively. So for each face image, we get 5×8=40 Gabor transformed images. So dimension reduction becomes necessary for the high dimensional extracted important facial features. In this paper independent component analysis (ICA) is done using FastICA. Thus some preprocessing becomes necessary on the extracted feature vectors before applying the FastICA [30]: i) Centering: Here the mean vector m = $E\{x\}$ is subtracted from x to make x a zero-mean variable. This preprocessing is made solely to simplify the ICA algorithms. As the face features have large difference mean, so centering is needed before implementation of FastICA. ii) Whitening: Whitening is done by $\tilde{x} = ED^{-1/2}E^T x$ where the diagonal matrix D and the orthogonal matrix E are obtained by PCA of xx^T . Whitening is done because whitening reduces the number of parameters to be estimated.

4.2 Implementation of FastICA

ICA learns the higher-order dependencies in addition to the second-order dependencies among the pixels. PCA driven coding schemes are optimal and useful only with respect to data compression and decorrelation of lower (second) order statistics. The ICA method, which expands on PCA as it considers higher (>2) order statistics, is used here to derive independent high intensity features found useful for the recognition of human faces. Whitening reduces the number of parameters to be estimated. So instead of estimating n^2 parameters that are the elements of the original matrix **A**, we need to estimate, only the mixing matrix **A** [30]. An orthogonal matrix contains n(n−1)/2 degrees of freedom. Whitening is done to reduce the complexity of the problem i.e. to reduce the dimension of the data. This is done by discarding the eigen values d_j of $E\{\mathbf{xx}^T\}$ which are too small.[31]. This also reduces the effect of noise. Again, dimension reduction also prevents over-learning. In this paper, the data has been preprocessed by centering and whitening, before using FastICA. Here the FastICA algorithm has been chosen based on a fixed-point iteration scheme proposed by Hyvärinen and Oja [31], which was derived using an approximative Newton iteration. The algorithm is motivated as follows:

1. Choose an initial random weight vector w.
2. Let $\mathrm{w}^{+} = \mathrm{E}\left\{\mathrm{x}g\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)\right\} - \mathrm{E}\left\{g'\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)\right\}\mathrm{w}$
3. Let $w = \mathrm{w}^{+} / \left\| \mathrm{w}^{+} \right\|$
4. If not converged, go back to 2.

The function $g(u)$ chosen here is the derivative of $f(u) = u^4$. Here it is here assumed that the data is prewhitened. The algorithm FastICA was introduced in two versions: i) a one-unit approach, and ii) a symmetric one. The first step, which is common for both versions and for many other ICA algorithms, consists in removing the sample mean the decorrelation of the data X, i.e., $\mathrm{Z} = \hat{\mathrm{C}}^{1/2}\left(\mathrm{X} - \overline{\mathrm{X}}\right)$ where $\hat{C} = \left(X - \overline{X}\right)\left(X - \overline{X}\right)^T$ and $\overline{X}$ is the sample mean of the measured data. The derivation of FastICA is as follows. The maxima of the approximation of the negentropy of $w^T x$ are obtained at certain optima of $\mathrm{E}\left\{\mathrm{G}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)\right\}$. According to the Kuhn-Tucker conditions (shown by Hyvärinen, A. and Oja, E. (1997), [31]), the optimum of $\mathrm{E}\left\{\mathrm{G}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)\right\}$ under the given constraint $\mathrm{E}\left\{\mathrm{G}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)^2\right\} = \left\|\mathrm{w}\right\|^2 = 1$ can be obtained at points where

$$E\left\{g\left(w^T x\right)\right\} - \beta w = 0 \qquad (6)$$

.Newton’s method can be used to solve equation (6). The Jacobian matrix of the equation (6) can be written as $JF(w) = E\left\{\mathrm{xx}^{\mathrm{T}} g'(w^T x)\right\} - \beta I$.

Since the data is sphere, it can be approximated in the following way:

$$\mathrm{E}\left\{\mathrm{xx}^{\mathrm{T}} g'(\mathrm{w}^{\mathrm{T}}\mathrm{x})\right\} \approx \left(\mathrm{xx}^{\mathrm{T}}\right)\mathrm{E}\left(g'(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right) = \mathrm{E}\left(g'(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right)\mathrm{I}$$

Thus the Jacobian matrix becomes diagonal, and can easily be inverted. Thus the following can be obtained with the approximative Newton iteration:

$$w^{+} = w - \left[\mathrm{E}\left\{\mathrm{x}g\left(\mathrm{w}^{\mathrm{T}}x\right)\right\} - \beta\mathrm{w}\right] / \left[\mathrm{E}\left\{g'\left(\mathrm{w}^{\mathrm{T}}x\right)\right\} - \beta\right] \qquad (7)$$

Usually the expectation of FastICA is replaced by their estimates. The natural estimates are generally the sample means. The one-unit ICA is based on minimization/maximization of the criterion $c(w) = E\left[G\left(x^T Z\right) - G_0\right]^2$ where w is the to-be found vector of coefficients that separates a desired independent components(ICs) from the mixture, E stands for the for the sample mean, $G(\cdot)$ is a suitable nonlinear function, called contrast function and G_0 is the expected value of $G(\eta)$ where η is a standard normal random variable. The symmetric FastICA estimates all signals in parallel, and each step is completed by a symmetric orthogonalization:

$$W^{+} \leftarrow g\left(\mathrm{WZ}\right)Z^{T} - \mathrm{diag}\left[g'\left(\mathrm{WZ}\right)1_N\right]W$$

$$W \leftarrow \left(\mathrm{W}^{+}W^{+T}\right)^{-1/2}W^{+},$$

where $g(\cdot)$ and $g'(\cdot)$,denote the first and the second derivative of $G(\cdot)$, respectively. Hyvärinen [31], formulated the likelihood in the noise-free ICA model, and then estimate the model by a maximum likelihood method, which is denoted by W= $(w_1,.., w_n)^T$ the matrix A^{-1}, the log-likelihood takes the form

$$\mathrm{W}^{+} = \mathrm{W} + \mu\left[1 + g(\mathrm{y})\mathrm{y}^{\mathrm{T}}\right]\mathrm{W} \qquad (8).$$

where μ is the learning rate.

Here, *g* is a function of the probability density function (pdf) of the independent components: $g = f_i' / f_i$ where f_{i} is the pdf of an independent component. The column vectors of W^{+} in (7) define a unique ICA representation. The feature base-vector of ICA is statistically independent. Thus it reflects the global feature of image and also local and edge features. Thus nature of image statistics can be more fully revealed by ICA. The IHIF method applies ICA on the high-intensity vectors obtained from (6). Lastly the IHIF method derives the overall ICA transformation matrix denoted by W^{+}. Next classification of the face images is done by the nearest neighbor algorithm, using different classifiers.

5. Experiment

The Gabor-based ICA method integrates the Gabor wavelet representation of face images and ICA for face recognition. When a face image is presented to the Gabor-based ICA classifier, the low-dimensional high intensity Gabor feature vector of the image is first calculated as detailed in Section 2, then fastICA is applied on the low-dimensional most important high intensity Gabor feature vector to finally obtain the most important independent high-intensity feature vector, which is used as the input data instead of the whole image. Thus we have input representations that are robust to noise and better reflect the data.

Let M_k^0 , k=1, 2…, l, be the mean of the training samples for class ω_k . The classifier applies, then, the nearest neighbor (to the mean) rule for classification using some similarity measure δ :

$$\delta\left(\Im, M_k^0\right) = \min_j \delta\left(\Im, M_k^0\right) \rightarrow \Im \in \omega_k \quad (9).$$

The independent high-intensity feature vector, $\Im$ is classified as belonging to the class of the closest mean, M_k^0, using the similarity measure δ. The similarity measure used here are, L_2 distance measure, δ_{L_2} , and the cosine similarity measure, $\delta_{\cos}$, which are defined as:

$$\delta_{L_2} = (X - Y)^T (X - Y), \quad (10)$$

$$\delta_{\cos} = \frac{-X^T Y}{\|X\|\|Y\|}, \quad (11)$$

where $\|\cdot\|$ denotes the norm operator, T denotes the transpose operator. Note that the cosine similarity measure includes a minus sign in (11) because the nearest neighbor (to the mean) rule of (9) applies minimum (distance) measure rather than maximum similarity measure.

5.1 Experiments of the Proposed method on Frontal and Pose-Angled Images for Face Recognition

This section assesses the performance of the Gabor-based ICA method for both frontal and pose-angled face recognition. The effectiveness of the Gabor-based ICA method is successfully tested on the ORL, FRAV2D and FERET face database. For frontal face recognition, the data set is taken from the ORL and the FRAV2D database. The dataset from ORL contains 400 frontal face images corresponding to 40 individuals, and the dataset from FRAV2D contains 1440 frontal face images corresponding to 120 individuals. The images are acquired under variable illumination, with occluded face features and facial expression. For pose-angled face recognition, the data set is taken from the FERET database, and it contains 1200 images with different facial expressions and poses of 120 individuals. Comparative performance of the proposed method is shown against the PCA method, the kernel PCA method. Performance results on FERET database has shown that proposed method is successive than the PCA, LDA and kernel PCA based algorithms.

5.1.1 ORL Database

The ORL database, employed here consists of 10 frontal face images of each individual consisting a total of 40 individual with different facial expressions and head pose (tilting and rotation up to 20 degrees), illumination condition also has slightly changes. Here each image is scaled to 92 × 112 with 256 gray levels. Fig. 3 shows all samples of one individual. Five images are randomly chosen from each individual of the 40 individual for training samples and the rest 5 images are used as testing samples.

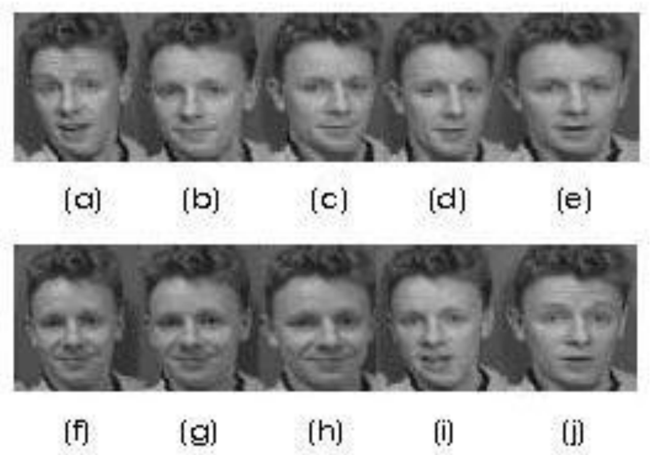


Fig. 3. Demonstration images of one individual from the ORL Database

5.1.2 FRAV2D Face Database

The FRAV2D face database, employed in the experiment consists; 1320 colour face images of 120 individuals, 12 images of each individual are taken, including frontal views of faces with different facial expressions, under different lighting conditions. All color images are transformed into gray images and are scaled to 92×112. Randomly five images of each individual were chosen i.e., is a total of 600 images as training samples and the remaining 840 images are regarded as testing samples. Fig. 4 show all samples of one individual.

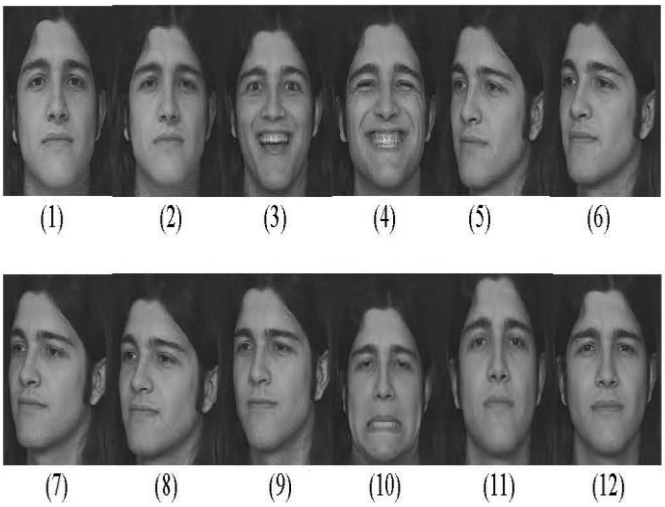


Fig. 4. Demonstration images of one individual from the FRAV2D database

5.1.3 FERET Face Database

The FERET database, employed in the experiment here contains, 1,200 facial images corresponding to 120 individuals with each individual contributing 10 images. The images in this database were captured under various illuminations and display a variety of facial expressions and poses. As the images include the background and the body chest region, so each image is manually cropped to exclude those, and finally scaled to 92×112. Fig. 5 shows all samples of one individual. Randomly five images from each individual were chosen i.e., is a total of 600 images are considered as training samples and the remaining 600 images are regarded as testing samples.

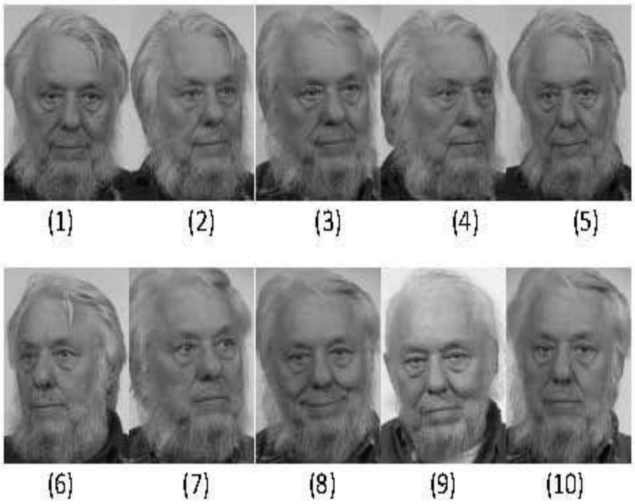


Fig. 5. Demonstration images of one individual from the FERET database.

5.2 Specificity and Sensitivity

Sensitivity and specificity [32] of the performance of the proposed method are measured on the dataset from the FERET, FRAV2D and ORL databases. In this regard, the true positive rate T_P; false positive rate F_P; true negative rate T_N; and false negative rate F_N: are also being measured. To measure the sensitivity and specificity, the dataset from the FERET database consists of a total of 100 class, in which each class have a total of 15 images, of which 10 images are of a particular individual, (in which 5 images are randomly used for training, and the left 5 images of the particular individual are used for positive testing), and the rest 5 images of other individuals are considered for negative testing. Fig. 6 shows all sample images of one class of the data set used from FERET database.

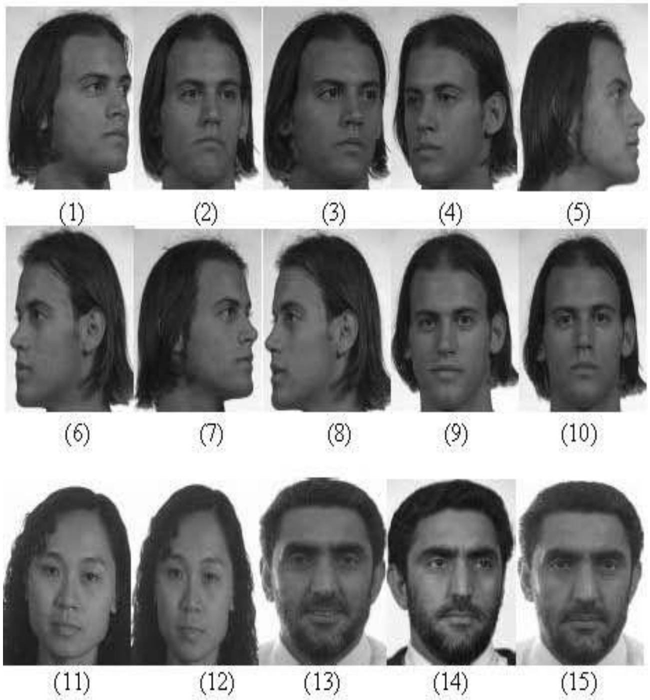


Fig. 6. Demonstration images of one class from the FERET database.

The dataset taken from the FRAV2D database consists of 120 classes, in which each class contains a total of 18 images, out of which 12 images are of particular individual, (in which 5 images are randomly used for training, remaining 7 images of the particular individual were used for positive testing), and the rest 6 images of other individuals were considered for negative testing. Fig. 7 shows all sample images of one class of the dataset used from FRAV2D database used to measure the sensitivity and specificity.

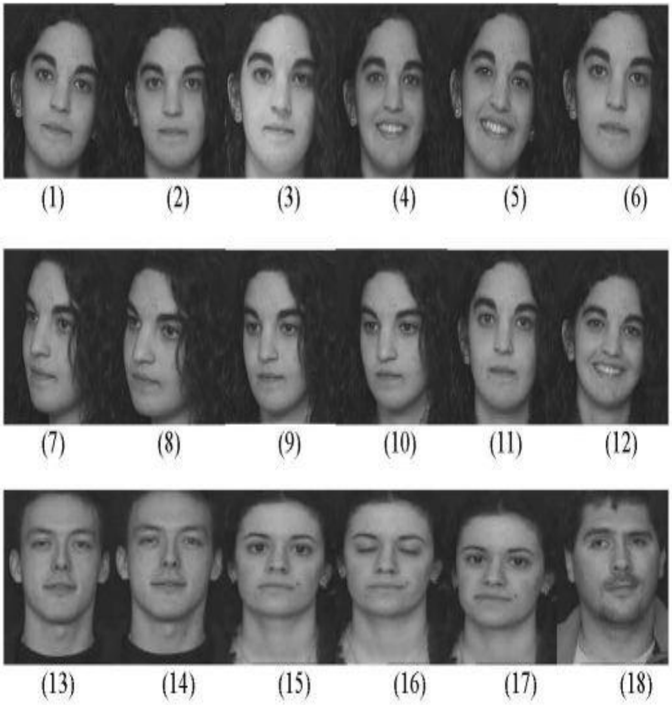


Fig. 7. Demonstration images of one class from the FRAV2D database

The dataset taken from the ORL database consists of 40 classes, in which each class contains a total of 15 images, out of which 10 images are of a particular individual, (in which

5 images are randomly used for training, remaining 5 were used for positive testing), and the remaining 5 images of other individuals are taken for negative testing. Fig. 8 shows all sample images of one class of the dataset used from ORL database used to measure the sensitivity and specificity.

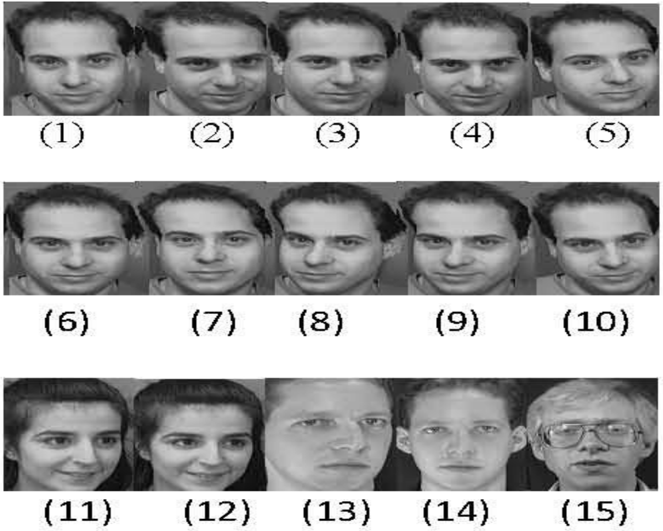


Fig. 8. Demonstration images of one class from the ORL database.

5.3 Experimental Results

Experiments were conducted that implements Gabor based ICA method, with different similarity measures on the ORL, FERET and FRAV2D database to measure both the positive and negative recognition rates i.e. T_P; F_P; T_N; F_N and hence the sensitivity and specificity. In this paper, we use only the high intensity features of the Gabor transformed image to include in ICA defined by (8). In order to derive the independent high-intensity ECA features. Fig. 9 and Fig. 10 shows face recognition performance of the IHIF method in terms of specificity and sensitivity respectively using the two different similarity measures. In our experiment, the algorithm would attempt to separate 500 ICs. Although it is shown already [12] that performance improves with the number of components separated. But, 1000 becomes a little bit intractable, so in order to have control over the number of ICs extracted by the algorithm, instead of performing ICA on the original images, ICA was applied on the set of extracted high-intensity feature vector obtained from the GW responses of the image. In Fig. 9 the horizontal axis indicates the number of ICs used, and the vertical axis represents the specificity rate of the face recognition, thus measures the proportion of negatives which are correctly identified which is the true negative rate i.e. correct rejection. The top response is the correct rejection rate. In Fig. 10 the horizontal axis indicates the number of ICs used, and the vertical axis represents the sensitivity rate of the face recognition, thus measures the proportion of actual positives which are correctly identified as such. From both the measure it is seen that the cosine similarity distance measure performs the best. This shows that cosine similarity distance measure further enhances face recognition.

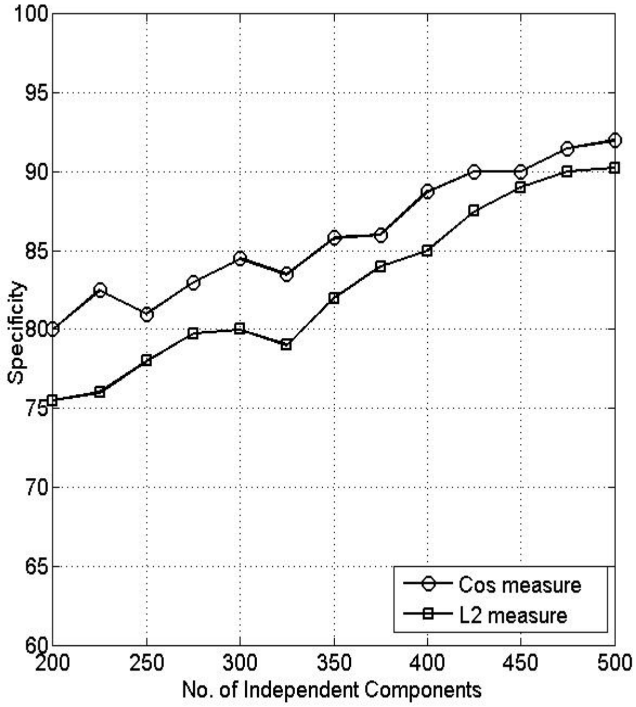


Fig. 9. Face recognition performance of the IHIF method on the FERET database, using the L_2 (the L_2 distance measure), and cos (the cosine similarity measure) for measuring negative recognition accuracy.

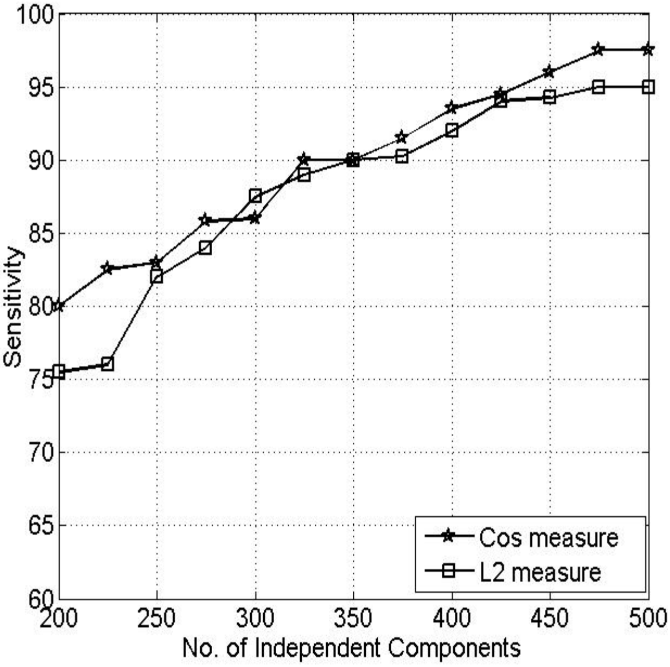


Fig.10. Recognition performance of the IHIF method on the FERET database, using the similarity measures L_2 (the L_2 distance measure), and cos (the cosine similarity measure) for measuring positive recognition accuracy.

As the proposed method performs best with the cosine similarity distance classifier, so experiments were conducted on the ORL, FERET and FRAV2D database to assess the face recognition performance in terms of specificity and sensitivity of the IHIF method, using the cosine similarity distance measure are shown in Fig.11 and Fig. 12.

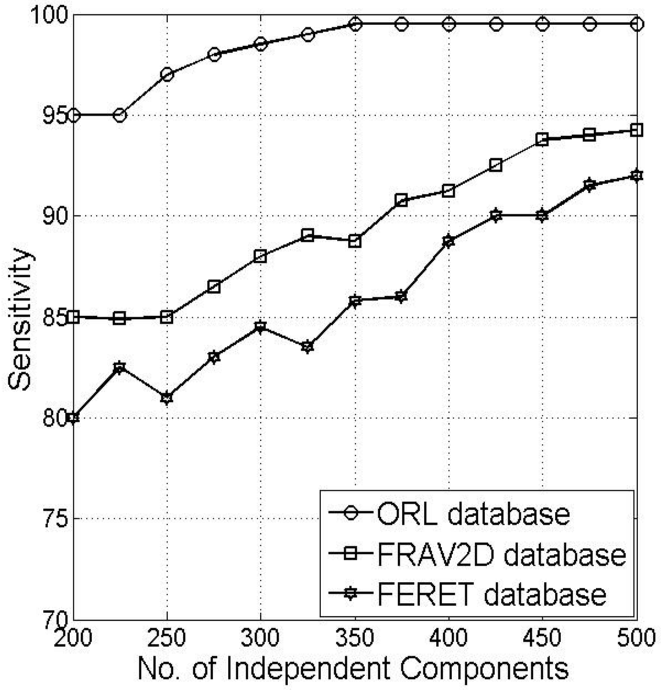


Fig. 11. Positive recognition performance of the IHIF method using the cosine similarity distance measure on the ORL, the FRAV2D and the FERET database.

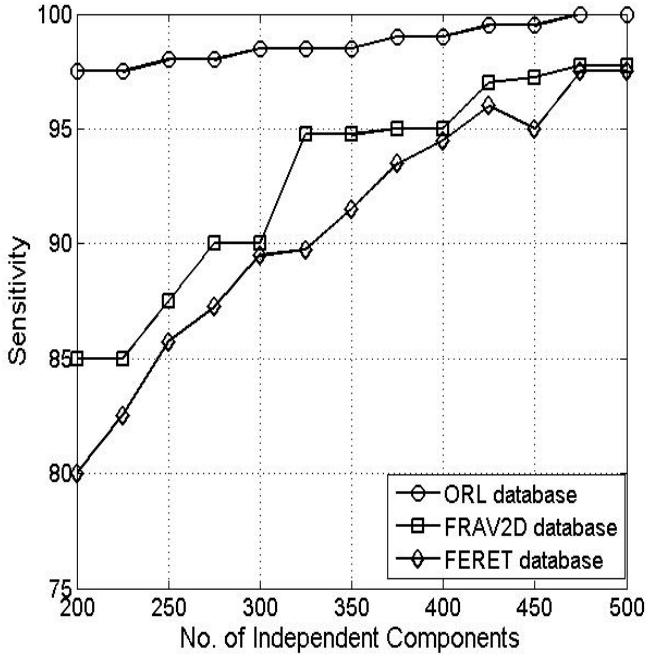


Fig. 12. Negative recognition performance of the proposed method using cosine similarity distance measure on the ORL,FRAV2D and the FERET database.

Specificity and Sensitivity measure of the proposed method for face recognition using the three databases, a) the FERET facial database, b) the FRAV2D facial database and the c) ORL database are shown in table1 table2 and table 3 respectively.

Table 1. The dataset from FERET database consists of 1200 images of 120 individuals,(in which each class onsists of 10 images of one individual and 5 images of any other individual/individuals) taken by permutation:

		Individual belonging to a particular class	
		Positive	Negative
FERET test	Positive	$T_P = 552$	$F_P = 12$
	Negative	$F_N = 48$	$T_N = 488$
		Sensitivity = $T_P / (T_P + F_N)$ = 92%	Specificity = $T_N / (F_P + T_N)$ ≈ 97.6%

False positive rate=$F_P/[F_P + T_N]$ = 1 − Specificity =2.9%
False negative rate = $F_N / (T_P + F_N)$ = 1 − Sensitivity= 8.0%
Accuracy = $(T_P+T_N)/(T_P+T_N+F_P+F_N)$ =94.8

Table 2: The data set from FRAV2D database consists of

1200 images of 100 individuals (in which each class folder consists of 12 images of one individual and 6 of any other individual/individuals) taken by permutation:

		Individual belonging to a particular class	
		Positive	Negative
FRAV2D test	Positive	(T_P) = 797	(F_P) =14
	Negative	(F_N)= 43	(T_N) =586
		Sensitivity = $T_P / (T_P + F_N)$ ≈ 94.85%	Specificity =$T_N / (F_P + T_N)$ =97.8%

False positive rate = $F_P / (F_P + T_N)$ = 1 − Specificity =**2.2%**
False negative rate = $F_N / (T_P + F_N)$ = 1 −Sensitivity=**5.15%**
Accuracy = $(T_P+T_N)/(T_P+T_N+F_P+F_N)$ ≈**96.32**.

Table 3: The dataset from ORL database consists of 400 images of 40 individuals (in which each class consists of 10 images of one individual and 5 image of any other individual/

/ individuals) taken by permutation:

		Individual belonging to a particular class	
		Positive	Negative
ORL test	Positive	(T_P) = 199	(F_P) =0
	Negative	(F_N)= 01	(T_N) =200
		Sensitivity = $T_P / (T_P + F_N)$ ≈ 99.5%	Specificity =$T_N / (F_P + T_N)$ = 100%

False positive rate = $F_P / (F_P + T_N)$ = 1 − Specificity =**0%**
False negative rate = $F_N / (T_P + F_N)$ = 1 − Sensitivity=**.5%**

Accuracy = $(T_P+T_N)/(T_P+T_N+F_P+F_N)$=**99.75.**

Recognition performances on ORL database of well-known methods are cited in previous works as: Eigenfaces 80.0% [33], Elastic matching 80.0% [33]. Neural network 96.2% [34], line based 97.7% min [35]. Although, reported recognition rates are higher in convolutional neural network and line-based method. It must be noted that these two are using more than one facial image for each individual in training. Our method achieves using only one reference facial image for each individual. So, the proposed method **gets** better results compared to similar methods. Tests on FERET database are held in accordance with the FERET procedure [36], and shown in table 4.

Table 4. Performance results of well-known algorithms and IHIF method on FERET database.

Method	**Recognition Rate (%)**
Eigenface method with Bayesian Similarity measure	79.0
Elastic graph matching	84.0
A Linear Discriminant Analysis based algorithm	88.0
Kernel PCA	91.0
Line based	92.7
Neural network	93.5
Proposed IHIF Method with Cos measure	**92**

The images considered here, consists of frontal faces with different facial expressions, illumination conditions, and occlusions. To examine the robustness of the proposed method to the illumination changes, the IHIF method has been experimented especially on FERET database. Experimental results indicate that a) the extracted high-intensity feature points by the proposed method enhance the face recognition performance in presence of occlusions as well as reduce computational complexity compared to the EGM [6]. b) The ICA applied on the extracted high-intensity feature vectors further enhances recognition performance. Our results show that 1) the IHIF method greatly enhance the recognition performance, and also reduces the dimensionality of the feature space when compared only with high-intensity features. 2) The cosine similarity measure classifier further enhances the recognition performance

6. Conclusion

In this paper a new approach to face recognition has been proposed by selecting the high-intensity features from the GWT facial images. Then ICA is applied on these extracted high-intensity feature vectors to obtain the IHIF vectors. As ICA employ local salient information, so here the extracted facial features are compared locally instead of a general structure, hence it allows us to make a decision from the different parts of a face, and thus maximizes the benefit of applying the idea of "recognition by parts". Thus it performs well in presence of occlusions (sunglasses scarf etc.), this is due to the fact that when there are sunglasses or any other obstacles the algorithm compares face in terms of mouth, nose and other features rather than the eyes. The algorithm also reduces computational cost as ICA is applied, only on the few extracted high-intensity features vectors of the GWT image instead of the whole image. Moreover, the proposed method has a simple matching procedure, and is also robust to illumination changes, as a property of GW's, together with the application of ICA. From the experimental results it is seen that proposed method achieves better results compared to other well known algorithms [33], [34], [35], [36] which are known to be the most successful algorithms. Experimental results also reveal that, ICA performs better using the cosines similarity distance measure than the Euclidean distance (L_2) measure.

Acknowledgement

Authors are thankful to the "Centre for Microprocessor Application for Training Education and Research" at the Department of Computer Science and Engineering, Jadavpur University, Kolkata - 700 032 for providing the necessary facilities for carrying out this work. Dr. D.K. Basu acknowledges with thanks the receipts of AICTE Emeritus Fellowship (1-51/RID/EF(13)/2007-08).

References

[1] M. H. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," in *Proc. IEEE Int. Conf. Img. Proc.*, Sept. 2000.

[2] P. Comon, "Independent component analysis, a new concept?" *Signal Proc.*, vol. 36, pp. 287–314, 1994.

[3] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Trans. Neural Networks*, vol. 8, pp. 486–504, Mar. 1997.

[4] Hyvärinen, A. and Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Computation*, 9(7):1483–1492.

[5] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Networks*, vol. 13, pp. 1450–1464, Nov. 2002.

[6] R. Chellappa, C. L.Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, pp. 705–740, May 1995.

[7] M.Turk and A.Pentland, "Eigenfaces for recognition." *Journal of Cognitive Science*, pp.71-86, 1991.

[8] P. Foldiak, "Forming sparse representations by local anti-Hebbian learning," *Biol. Cybern.*, vol. 4, pp. 165–170, 1990.

[9] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1357–1362, Dec. 1999.

[10] J.P. Nadal and N. Parga, "Non-linear neurons in the low noise limit: A factorial code maximizes information transfer," *Network*, vol. 5, pp. 565–581, 1994.

[11] P. Hancock, "Alternative representations for faces," in *British Psych. Soc., Cognitive Section*. Essex, U.K.: Univ. Essex, 2000.

[12] M.S. Barlett, G. L. Donato, J. R. Movellan, J. C Hager P. Ekman, and T. J. Sejnowski, "Img. representations for facial exp. coding," in *Adv. in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Muller, Eds. Cambridge, MA: MIT Press, 2000, vol.12.

[13] M. Lades, J Vorbruggen, J, Buhmann, J. Lange, von der Malsburg, and R. Wurtz, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, no. 3, pp. 300-311, 1993.

[14] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[15] Burcu Kepenekci , F. Boray Tek . Gozde Bozdagi Akar, "Occluded face recognition based on Gabor Wavelets," IEEE, ICIP, vol. 6, pp. 293-296, Feb. 2002.

[16] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 13, pp. 607–609, 1996.

[17] Arindam Kar, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu and Mahantapas Kundu, "Classification Of High-Intensity Gabor Responses Using Bayesian PCA For Human Face Recognition", IJRTE Journal Nov 2009 issue pp. 106-110.

[18] Zhang qiang, Chen chen, Zhou changjun*, Wei xiaopeng, "Independent Component Analysis of Gabor Features for Facial Expression Recognition", 2008 International Symposium on Information Science and Engineering,pp. 84-87.

[19] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[20] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection" *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.

[21] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 349–361, Apr. 2001.

[22] The Face recognition technology (FERET) Database, http://www.itl.nist.gov/iad/humanid/feret.

[23] Face Recognition and Artificial Vision group FRAV2D face database http://www.frav.es/

[24] Olivetti & Oracle Research Laboratory, The Olivetti & Oracle Research Laboratory Face Database of Faces, http://www.cam-orl.co.uk/facedatabase.html.

[25] L. Wiskott, J. M. Fellous, N. Krüger and Christoph von der Malsburg,"Face Recognition by Elastic Graph Matching," *In Intelligent Biometric Techniques in fingerprint and Face Recognition,* CRC Press, Chapter 11, pp. 355-396, 1999.

[26] J. G. Daugman, "Two dimensional spectral analysis of cortical receptive field profile", *Vision Research*, vol. 20, pp. 847-856, 1980.

[27] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, pp. 429-459, 1946.

[28] C. Liu and H. Wechsler, "Independent Component Analysis of Gabor Features for Face Recognition," IEEE Trans. Neural Networks, vol. 14, no. 4, pp. 919-928, 2003.

[29] D. H. Hubel and T. N. Wiesel, "Functional arch. of macaque monkey visual cortex," *Proc. Royal Soc. B (Lond)*, vol. 198, pp.1-59, 1978.

[30] T.W. Lee, M. Girolami, and T. J. Sejnowski, "Indp. comp. analy. using an extended infomax alg. for mixed sub-Gaussian & super-Gaussian sources," *Neural Comput.*, vol. 11, no. 2, pp. 417–41, 1999.

[31] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Trans. Neural Networks*, vol. 8, pp. 486–504, Mar. 1997.

[32] http://en.wikipedia.org/wiki/Specificity_and_sensitivity

[33] J. Zhang, Y. Yan. and M. Lades, "Face Recognition: Eigenface. Elastic Matching. and Neural Nets," *Proc.IEEE,* vol. 85**,** pp.1423- 435. 1997.

[34] ***S.*** Lawrence, C. Giles, A. Tsoi. and A. Back. "Face Recognition: A Convolutional Neural Network Approach," *IEEE Trans. on Neural Networks,* vol. 8. pp. 98-113, 1997.

[35] 0. de Vel and S. Aeberhard. "Line-Based Face Recognition Under Varying Pose." *IEEE Trans. on PAMI.* vol. 21. no. 10, October 1999.

[36] P. Phillips, "The FERET database and evaluation procedure for face Recognition algorithms",. Image and Vision Computing, vol. 16, no. 5, pp. 295-306, 1998.

A New Protocol for Advanced Encryption/Decryption Using ECC & PHAL

Yogendra Kumar Jain[1], Nitin[2]
[1]Head of the Computer Science & Engineering, Samrat Ashok Technological Institute; Vidisha, M.P., India,
E-mail: ykjain_p@yahoo.co.in
[2]Research Scholar, Computer Science & Engineering, Samrat Ashok Technological Institute Vidisha, M.P., India,
E-mail: nitin4u.in.sati@gmail.com

Abstract- *The biggest issue for the data over internet world is the security of information is the. The meaning of security is to protect the information from the unauthorized person or hackers. In general, the data security cryptography techniques are RSA, DSA and hashing algorithm such as SHA, MD5 and RIPEMD-160. The Parameterized Hash Algorithms (PHAL) was proposed dedicated hash algorithm and designed as an answer for weaknesses of MD/SHA hash function family. The attacks on well-known and widely used hash functions motivate us to design of a new security technique and main contribution of proposed work is to provide security for data over internet. In this paper, we proposed a hybrid of ECC and PHAL algorithms for encryption and decryption. The public key and the private key of ECC are used in PHAL algorithm. When ECC is hybrid with PHAL, the strength of security is improved significantly. Such combination takes very less time in encryption and decryption in comparison of RSA-SHA combination. Therefore, the proposed method improves the strength of encryption as well as improves the speed of processing. By using the proposed method, the time attack can be prevented effectively in chipper data (data that has no meaning) and avoiding the danger of eavesdropping of private keys. This scheme is effective, secure, and easy to implement for wired/wireless networks. All the mechanisms thoroughly discussed in this paper, proved to work well together and provide the needed security in any environment.*

Keywords: *Network design, Security, ECC, PHAL, Encryption, Decryption.*

1. Introduction

The evolution of wireless networking has raised some most unique and compelling issues. The important and biggest issue is security. Security becomes very critical with the increasing possible numbers of attacks. Electronic Commerce (EC) has been expanding rapidly in quality and quantity since it started on the internet. The reason is that, it can be done by increasing the reliability of EC with the new development of security technique. The SSL, a Security Socket Layer, which is currently used in EC is being considered the only stable access to internet during the transportation, but it can hardly ensure the problem of information security. Information security is concerned with the confidentiality, integrity and availability of data regardless of the form the data may take: electronic, print, or other forms [1]. Such a protocol is related directly to cryptography for security and consists of an asymmetric key algorithm, RSA for authentication and non-repudiation, DES for the message confidentiality, Hash algorithm and SHA for message integrity. But the disadvantage of the suggested protocol is that the speed of processing is slow because of long key size. From the standpoint of this, elliptic curve cryptosystem (ECC) technique is very important for cryptography. In this paper, we propose a new method for encryption and decryption of data, which uses hybrid of ECC and PHAL [2]. The proposed method improves the strength of encryption as well as improves the speed of processing. The public key and the private key of ECC are used in the PHAL algorithm. For PHAL hash algorithm, a similar approach was chosen. Additionally, the number of *rounds* was added as a parameter [3]. The design goals of PHAL algorithm are determined as follows:

- Hash size should be 256 bits of length.
- Its iteration structure should have resistant against known attacks against the MD-type structure.
- Its structure should have resistant against the known attacks.
- Its structure should have parameterized, to reach flexibility between performance and security.

Therefore, the digital envelope used in the existing ASEP protocol can be removed by the ECC and PHAL algorithm, which simplifying the complexity of dual signature. The rest of the paper is organized as follows. In section 2, literature review of previously related theory, in section 3, cryptanalysis and security related issues are described and section 4 presents the proposed algorithm using hybrid of ECC and PHAL algorithms. Section 5 shows the experimental results and performance comparison of proposed method with RSA, ECC and, SHA-RSA.

2. Literature Review

In the recent years, vigorous research has been carried out in analysis of security, encryption and decryption. Although, different technologies have been devised for the information security, especially for online information security, but still the area is being explored. The different approaches have been proposed by the various researchers on the basis of features extracted.

In 1988, Chaum et. al. proposed a protocol which is relied on a single use token method. They presented that the user creates blinded e-bank currency note and passes it to the bank to be signed using bank public key. The bank signs the currency note, subtracts the value from the user account, and returns the signed currency note back to the user. The user removes the blind thing and utilizes it to buy goods from the super market. The super market checks the authenticity of the bank currency note using the bank public key and passes it to the bank where they are verified contrary to a list of currency note already used. The amount is deposited into the supermarket account, the deposit approved, and the supermarket in turn emits the merchandise [4].

In the year 1995, Glassman et. al. presented a decentralized e-payment protocol. They employed a type of e-coins and utilize asymmetric encryption techniques for all information transactions. Millicent is a lightweight and secure scheme for e-commerce through the internet. It is developed to support to buy goods charging less than a cent. It is relied on decentralized validation of e-currency at the seller server without any further communication, costly encryption, or off-line processing [5].

In the year 1997, Rivest suggested that there is a possibility to reduce the number of messages engaged with every transaction. Also, the lottery ticket scheme is relied on the assumption that financial agents are risk neutral and will be satisfied with fair wagers [6]. While in 1998, Foo et. al. proposed a payment scheme using vouchers [7]. The e-vouchers can be moveable but the direct exchange between purchasers and vendors is impossible. As a result, a financial agent is needed and this will raise the transactions charges of exchange.

Kim and Lee proposed an efficient and flexible protocol [8] in 2003. The presented scheme supports multiple merchant payments and prevents overspending payment. Moreover, in pay-word system, whenever a customer wants to establish transactions with each vendor, he has to obtain a certificate from a broker and create a series of pay-words, while a customer is able to make transactions with different merchants by performing only one hash chain operation.

In 2004, Lee et. al. [9], proposed an ASEP (Advanced Secure Electronic Payment) protocol. They uses ECC, SHA (Secure Hash Algorithm) and 3BC (Block Byte Bit Cipher) instead of RSA and DES. They shown that ASEP protocol has an simple structure and improve the performance with the length of session key, byte-exchange algorithm, bit operation algorithm, and so on. From the standpoint of the supply for key, the CA (Certificate authority) has only to certify any elliptic curve and any prime number for modulo operation, the personally identifiable information and security for information can be guaranteed over insecure network.

In the year 2008, Chen-Lee proposed a novel mutual authentication protocol using hash function to solve remote user authentication. Which is continuing by the beneficial assumption of the Peyravian – Zunic schemes [10], and the server is not required to maintain a security-sensitive table. Their scheme is secure under assumption of well-defined tamper-resistant smart card device, which can prevent the hackers from reading the secret messages [11].

In 2009, Jian-zhi et. al. [12] transplanted the hyper elliptic curve (HEC) system into DSA algorithm. They use QuartusII6.0 and FPGA to implement the algorithm and generate the system Module. The QuartusII6.0 was used to synthesis the system and generates RTL and simulated waveform. The digital signature system combines HEC and DSA. It inherits their securities. It provides high security to check the integrality of file and ID distinguish. The presented method can solve the problem that how to check integrity of the file and signature ID, and it especially fit for the internet operation that need identity validate. This scheme is based on only digital signature algorithm.

In the year 2009, Subasree et. al. [13] proposed the new Public Key Cryptographic algorithm for better security with integrity using a combination of both symmetric and asymmetric cryptographic techniques. It is also called RSA-CRT, because it is used Chinese Remainder Theorem [14]. CRT is used for Decryption. It is shown that Dual-RSA improved the performance of RSA in terms of computation cost and memory storage requirements. It achieves parallelism, but this one is also taking more time for the encryption and decryption. The most remote user authentication schemes are designed by presented cryptographic techniques. In terms of computational cost, the hash-function-based scheme is more simple and efficient.

In the year 2009, Lin and Xie [15] define a method based on mutual authentication. This scheme uses hash function in the authentication phase. The authentication server needs not maintain a verifier table to verify the validity of the user login. The scheme can achieve mutual authentication. The user can change password over public networks. The scheme can resist all sorts of attacks, such as replay attack, password guessing attack, modification attack, impersonation attack, forgery attack, stolen attack and denial-of-service attack but not timing attack.

In 2009, Aboud has discuss an important e-payment protocol namely Kim and Lee scheme examine its advantages and delimitations, which encourages the author to develop more efficient scheme that keeping all characteristics intact without concession of the security robustness of the protocol [16]. He suggested a protocol employs the idea of public key encryption scheme using the thought of hash chain and compared the proposed protocol with Kim and Lee protocol and demonstrates that the proposed protocol offers more security and efficiency, which makes the protocol workable for real world services. The protocol is divided into three schemes: certificate issuing scheme, payment scheme, and redemption scheme. However, several e-payment protocols [17] [18] [19] have been suggested in the recent past years. Therefore, our proposed model is inspiring by the previously proposed technique. We will try to eliminate the existing timing attack problem.

3. Background Theory

3.1. Encryption and decryption algorithm

Our scheme employs some basic concepts, such as one-way hash function, e.g., MD5 [20], or SHA-1 [21], discrete logarithm problem [22], and Diffie–Hellman key agreement protocol [23]. The basic concepts have been described in the following subsections.

3.1.1. One-way hash function

In public key cryptography, keys and messages are expressed numerically and operations are expressed mathematically. The private and public key of a device is related through the mathematical function called the one-way function. One-way functions are mathematical functions in which the forward operation can be done very easily but the reverse operation is so difficult that it is practically impossible. In public key cryptography the public key is calculated by using private key on the forward operation of the one-way function.

A one-way hash function H: A→B is a function with the following properties:

- The one-way hash function H takes a message of arbitrary length as the input and produces message digest of fixed-length as the output.
- The one-way hash function H is one-way hash in the sense that given a, it is very easy to compute H (A) = B. However, given b, it is hard to compute H^{-1} (B) = A.
- Given A, it is computationally infeasible to find such that $A^{/} \neq A$, but H $(A^{/})$ = H (A).
- It is computationally infeasible to find any pair A, $A^{/}$ such that $A^{/} \neq A$, but H $(A^{/})$ = H (A).

3.1.2. Discrete logarithm problem

Solving discrete logarithm problem is still a hard problem. We describe this problem as follows. Assume that g is a generator of Z^*_p and p is a large prime number [22]. Z^* is the multiplicative group.

Consider the following equation:

$J = g^j \bmod p.$ (1)

If we know g, j, and p, it is very easy to compute J. However, if we know g, J, and p, it is very difficult to solve the equation for j. The difficulty is due to factoring prime numbers as that required for RSA [24]. The problem of solving Eq. (1) for j is called discrete logarithm problem.

3.1.3. Diffie–Hellman key agreement

In 1976, Diffie and Hellman proposed a key agreement scheme for making agreement on a session key over insecure communication networks [23]. The scheme allows two users communicate each other in an insecure communication network with the decided session key. Its security is based on solving discrete logarithm problem. Assume that Alice and Bob are to agree on a session key over insecure communication networks. The parameters g and p are public. Then, they do the following steps to agree on a session key.

- Alice randomly chooses a large number n and sends Bob $A = G^n \bmod P$.
- In the meantime, Bob also randomly chooses a large number m and sends Alice $B = G^m \bmod P$.
- After that, Alice and Bob can calculate their session key as $K = B^n \bmod P = A^m \bmod P = G^{nm} \bmod P$.

Without knowing n and m, no one can listen on the Alice–Bob channel. To derive n and m, it is discrete logarithm problem.

4. The Proposed Hybrid ECC & PHAL Technique

4.1. Description of PHAL (Parameterized Hash Algorithm)

In this section, we described Parameterized Hash Algorithm Design Strategy [2] was discovered in 2008 by P. Rodwald (*Military Communication Institute, Poland*) and J. Stokłosa (*Poznań University of Technology, Poland*). It's designed to be not only secure but also flexible.

4.1.1. Message Padding

The message (M) has to be padded before hash computation begins. The length of padded message should be a multiple of m bits. The message is padded by appending a zero or greater number of bits "0" until the length of the message is congruent to (m-72) mod m. Finally, we append 4-bit digest length d mod32, then 4-bit length value rounds which defined number of rounds and at the end appends original message length (mod 2^{64}). The message M is then divided into k m-bit length blocks $M_0, M_1, \ldots, M_{k-1}$.

4.1.2. Iteration Schema

The number of bits hashed so far (counter) was added to increase resistance of hash function to fixed points attacks. Random value (salt) increases resistance of hash function to attacks which use precomputation table generated in advance (message - hash value). Number of rounds (rounds) was added to make this function more flexible. There is a trade-off between performance and security. Small number of rounds should be used in systems where performance is most important. When security is the most important factor, greater number of rounds should be used. More security factors connected with parameters salt and counter can be found in HAIFA design analysis [3].

4.1.3. Initial Value

In many hash functions the first 32-bits of the fractional parts of the square roots of the first 8 prime numbers are taken as an initial value. In PHAL-256 as an initial value a balanced vector was chosen.

Hamming weight of each word is equal to 16, and Hamming weight of each bit position is equal to 4.

4.1.4. Message Modification

Each m-bit message block Mi is divided into sixteen *w*-bit words mi (0)… mi (15). Before each round r, except the first one 2 ≤ r ≤ round, words are modified three times using the following schema. Before the first round substitution $w_{\mathrm{i}}^{1} = m_{\mathrm{i}}$ must be done. Before the second round substitution $w_{\mathrm{i}}^{2} = w_{\mathrm{i}}^{1}$ must be done.
As a result modified message w_i^r is obtained. This message is then used in the round function and is used as an input for message modification for the next round. Both branches $BRANCH_b$, for $0 \le b \le 1$, use modified message words with different order σ_b. Each branch uses each message word twice in single round.

4.1.5. Modified Message Word Ordering

In addition to the fact that an original message is modified before computation, each branch has its own message order. This was done as an answer to Wang at al.'s attacks against RIPEMD family [25]. RIPEMD-128/160, due to different message-ordering in branches, is still not broken by their attacks. If an attacker constructs an intendant differential characteristic for one branch, the different word order in the second branch will cause unintended differential patterns. The order of message words was chosen with the conditions: balancedness in upper, lower, left and right part.

4.1.6. Constants

In many hash functions the first thirty-two bits of the fractional parts of the cube roots of the first 16/64 prime numbers are taken as constants. In PHAL-256 as constants twelve balanced words were chosen. Hamming weight of each word is equal to 16, and Hamming weight of each bit position is equal to 6. *S-boxes* Two S-boxes were generated with the following parameters: high nonlinearity (74), balancedness and good XOR profile.

4.1.7. One-argument Functions - g

Functions *g*1 and *g*2 output one word with one input word. All possible functions $g(x) = x \oplus (x\ n) \oplus (x\ m)$, for $n, m \in \{1... 31\}$ and all (2^{32}) possible values of input vector were investigated. Values of shift rotations were chosen from sets satisfying the following conditions:
1. If $HW(x) = 1$, then $HW\,(g(x)) \ge 2$,
2. If $HW(x) = 2$, then $HW\,(g(x)) \ge 4$,
3. If $HW(x) = 3$, then $HW\,(g(x)) \ge 3$,
4. *n* and *m* are not divisors of 32,
5. $4 < n < 28$, $4 < m < 28$,
6. $|n - m| > 8$,
7. If *n* is even, then *m* is odd,
8. If *m* is even, then *n* is odd,
Where, HW means Hamming Weight. By above conditions, functions *g*1 and *g*2 were defined.

4.1.8. Chaining Value

In the most popular hash functions many words of chaining variable are not modified in a single step. They are just copied. Additionally, output words of Boolean functions are used to update only one chaining variable. The situation in PHAL-256 is different. Each word of chaining variable is modified in a single step at least twice: once with message word and once with usage of function: *g*, *f* or *S*.

4.2. Elliptic Curves in Cryptography (ECC)

Elliptic Curve Cryptography (ECC) was discovered in 1985 by Victor Miller (IBM) and Neil Koblitz (University of Washington) as another mechanism for implementing public-key cryptography. Public-key algorithms create a mechanism for sharing keys among large numbers of parties in a complex information system. Unlike other popular algorithms such as RSA, ECC is based on discrete logarithms problem that are much more difficult to challenge at equivalent key lengths.

When using elliptic curves [26] in cryptography, we use various properties of the points on the curve, and functions on them as well. Thus, one common task to complete when using elliptic curves as an encryption tool is to find a way to turn information *n* into a point P on a curve E. We assume the information *n* is already written as a number. There are many ways to do this, as simple as setting the letters a = 0, b = 1, c = 2 . . . or there are other methods, such as ASCII, which accomplish the same task. Now, if we have E: $y^2 = x^3 + Ax + B \pmod{p}$, a curve in Weierstrass form, (The elliptic curve is given by the Weierstrass equation $y^2 + A_1xy + A_3y = x^3 + A_2x^2 + A_4x + A_6$. This is specialized with variable changes to the equations initially shown) we want to let $n = x$. But, this will only work if $n^3 + An + B$ is a square modulo p. Since only half of the numbers modulo p are squares, we have about only a half chance of this occurring. Then, we will try to implant the information *n* into a value that is a square. Pick some K such that $1/2^K$ is an acceptable failure rate for implanting the information into a point on the curve. Also, make sure that $(n + 1)\,K < p$. Let $x_j = nK + j$ for $j = 0, 1, 2. . . K - 1$. Compute $x^3j + Ax_j + B$. Calculate its square root $y_j \pmod{p}$, if possible. If there is a square root, we let our point on E representing m be $P_n = (x_j, y_j)$ If there is no square root, try the next value of j [27, 28].
So, for each value of j we have a probability of about 1/2 that x_j is a square modulo p. Thus, the probability that no x_j is a square is about $1/2^K$, which was the satisfactory failure rate [29]. In most common applications, there are many real-life problems that may occur to destroy an attempt at sending a message, like system or electricity failure. Since people accept a certain many amount of failure due to uncontrollable occurrence, it makes sense that they could agree on an acceptable rate of failure for a controllable feature of

the process. Though this we will not use that specific process in our algorithms [30].

As shown in Figure 1, the user A computes a new key K_A (K_BP) by multiplying the user B's public key by the user A's private key K_A. The user A encodes the message by using this key and then transmits this cipher text to user B. After receiving this cipher text, the user B decodes with the key K_B (K_AP), which is obtained by multiplying the user A's public key, K_AP by the user B's private key, K_B.

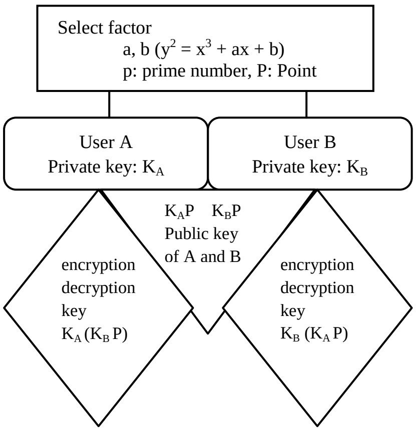


Figure: 1. Concept of en/decryption of ECC

Therefore, as K_A (K_BP) =K_B (K_AP), we may use these keys for the encryption and the decryption.

4.3. Proposed Method

In that section we described the working of our proposed method:

Suppose that Alice wants to send a signed message to Bob. Initially, the curve parameters (q, FR, a, b, [DomainParameterSeed], G, n, h) must be agreed upon. q is the field size; FR is an indication of the basis used; a and b are two field elements that defined the equation of the curve; [DomainParameterSeed] is a optional bit string that is present if the elliptic curve was randomly generated in a verifiable fashion; G is an base point of prime order on the curve (i.e., G = (x_G,y_G)); n is the order of the point G; and h is the cofactor (which is equal to the order of the curve divided by n).

Also, Alice must have a key pair suitable for elliptic curve cryptography, consisting of a private key K_A (a randomly selected integer in the interval [1, n − 1]) and a public key K_AP (where $K_AP = K_AG$). Let L_n be the bit length of the group order n.

For Alice to sign a message m, she follows these steps:

1. Calculate e = P.HASH (m), where P.HASH is a cryptographic parameterized hash function, such as PHAL, and let z be the L_n leftmost bits of e.
2. Select a random integer q from [1, n − 1].
3. Calculate r = x_1(mod n), where (x_1, y_1) = qG. If r = 0, go back to step 2.
4. Calculate s = $q^{-1}(z + rK_A)$ (mod n). If s = 0, go back to step 2.
5. The signature is the pair (r, s).

When computing s, the string z resulting from P.HASH (m) shall be converted to an integer. Note that z can be greater than n but no longer.

4.3.1. Verification Method

For Bob to authenticate Alice's signature, he must have a copy of her public key K_AP. If he does not trust the source of K_AP, he needs to validate the key (I; here indicate the identity element):

1. Check that K_AP is not equal to I and its coordinates are otherwise valid.
2. Check that K_AP lies on the curve.
3. Check that $nK_AP = I$.

After that, Bob follows these steps:

1. Verify that r and s are integers in [1, n − 1]. If not, the signature is invalid.
2. Calculate e = P.HASH (m), where PHAL is the same function used in the signature generation. Let z be the L_n leftmost bits of e.
3. Calculate w = s^{-1}(mod n).
4. Calculate u_1 = zw (mod n) and u_2 = rw(mod n).
5. Calculate (x_1, y_1) = $u_1G + u_2K_AP$.
6. The signature is valid if r = x_1(mod n), invalid otherwise.

Note that using Straus's algorithm (also known as Shamir's trick) a sum of two scalar multiplications $u_1G + u_2K_AP$ can be calculated faster than with two scalar multiplications.

5. Simulation Results, Analysis, and Performance Evaluation

Cryptography is used to achieve few goals like Confidentiality, Data integrity, Authentication etc. of the communicated data. In order to achieve these goals, various cryptographic algorithms have been developed by various people in the past years. To achieve these goals, we also proposed a new algorithm, which is composite of ECC and PHAL algorithm and provides better performance in comparison to the existing methods.

5.1. The Composition of ECC with PHAL

In this paper, we proposed the hybrid of ECC and PHAL, instead of RSA and SHA. ECC is faster than RSA for signing and decryption, but slower than RSA for signature verification and encryption. New iteration schema in PHAL has additionally provided few desirable properties: maintaining the collision resistance of the compressions function, increasing the security of iterative hash functions against pre-image attacks.

Table 1: A comparison for encryption time
Unit: µs

Method of encryption key size	RSA	ECC (F_2^m)	RSA & SHA	ECC & PHAL
5	0.05	0.05	0.0025	0.0012
10	0.54	0.20	0.0051	0.0025
15	1.34	0.29	0.0076	0.0038
20	2.55	0.38	0.0102	0.0051
25	4.33	0.42	0.0130	0.0064
50	5.53	0.85	0.0256	0.0128

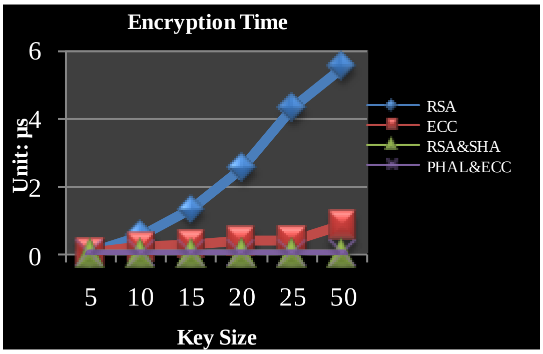
Table 2: A comparison for decryption time
Unit: µs

Method of decryption key size	RSA	ECC (F_2^m)	RSA & SHA	ECC & PHAL
5	0.11	0.03	0.0024	0.0012
10	0.55	0.04	0.0050	0.0025
15	1.20	0.04	0.0076	0.0038
20	3.08	0.04	0.0102	0.0051
25	6.21	0.05	0.0129	0.0064
50	8.06	0.05	0.0256	0.0129

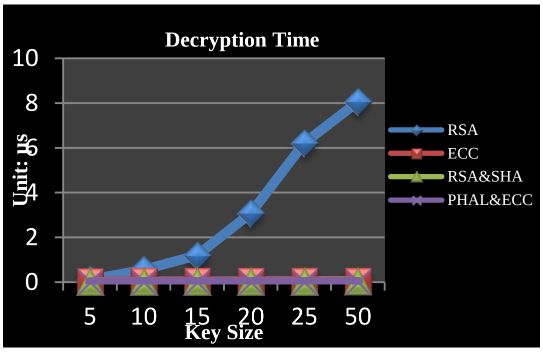
We will compare the performance of our proposed hybrid of ECC-PHAL with RSA, ECC and SHA. The results of the encryption and decryption times are shown in Tables 1 and 2 respectively, which indicates that encryption and decryption time of our proposed method is much less than those of RSA, ECC and SHA.

5.12 Experimental Results and Analysis

In this section, we will show that the proposed new method give better result. In our security technique, timing attack as well as brute-force attacks are not possible. ECC combined with PHAL Encryption makes it ideal for applications such as encrypting cell-phone calls, credit card transactions, and other applications where memory and speed are an issue. The encryption and decryption time is very less in comparison to the existing techniques. The graph 1 and 2 as shown below represents the comparison of encryption time and decryption time for our proposed method with RSA, ECC RSA-SHA.



Graph 1: Encryption time



Graph 2: Decryption time

The results show that the proposed method speeds up the encryption process by reducing communication traffic for transmission, and simplifying the dual signature. In addition, the security for information is strengthened which prevents session keys from being intercepted from attackers on the network.

Additionally, digital signature properties of this mechanism ensure the non-repudiation of origin for the whole messages exchange.

Data can be secured and protected against any outer theft and tampering, especially when data is being sent between branches, through the encryption and decryption using ECC and PHAL algorithm. Pre-shared keys are the simplest authentication method to implement and permit two branches communicate with each other in private, and their private key should exist the same and never given out.

6. Conclusion

This article proposed new protocol for advanced encryption / decryption using ECC & PHAL with the latest technology. The proposed data encryption / decryption technique employees ECC and PHAL algorithm instead of RSA and SHA used in the existing cryptography techniques. The results show that the proposed method speeds up the encryption process by reducing communication traffic for transmission, simplifying dual signature. In addition, the security for information is strengthened. PHAL family looks resistant against existing attacks, in particular against Wang at al.'s attacks [31]. Three rounds seem to be optimal trade-off between security and performance. The design shows how the network can be more secure by encrypting the sending data using combination of ECC and PHAL between user to server and server to user. The anonymity and security for information can be guaranteed over communication network. The purpose of network security is to provide availability, integrity, and confidentiality. Thus, the main objective of encryption and decryption is to prevent outsiders (hackers) from interfering with messages sent among hosts in the network, and to protect the privacy and integrity of messages going through untrusted.

References

[1]. P. Rodwald, P. Mroczkowski, "Iteration schema of Parameterized Hash Algorithm (PHAL)", Military Communications and Information Systems Conference – MCC 2007, Bonn, Germany, 2007

[2]. P. Rodwald, J. Stokłosa, "Family of Parameterized Hash Algorithms" The Second International Conference on Emerging Security Information, Systems and Technologies, IEEE, pp. 203-208, 2008.

[3]. D. Johnson, "Improving Hash Function Padding", Technical report, National Institute of Standards and Technology NIST, Oct. 2005.

[4]. D. Chaum, Fiat and M Naor, "Untraceable electronic cash", In Proceeding Advances in Cryptology, LNCS 403, Springer-Verlag, pp. 319-327, 1988.

[5]. S Glassman, M Manasse, M Abadi, P Gauthier and P Sobalvarro, "The Millicent protocol for inexpensive electronic commerce", In Proceeding of the International World Wide Web Conference, O'Reilly, pp. 603–618, 1995.

[6]. R Rivest, "Electronic lottery tickets as micropayments", In Proceeding of the International Conference of Financial Cryptography, LNCS 1318, Springer-Verlag, pp. 307–314, 1997

[7]. E Foo and C Boyd, "A payment scheme using vouchers", In Proceeding of the International Conference of Financial Cryptography, LNCS 1465, Springer-Verlag, pp. 103-121, 1998.

[8]. S Kim and W Lee, "A Pay-word-based micro-payment protocol supporting multiple payments", In Proceeding of the International Conference on Computer Communications and Networks, pp. 609-612, 2003.

[9]. Byung Kwan Lee, Tai-Chi Lee Seung Hae Yang "An ASEP (Advanced Secure Electronic Payment) Protocol Design Using 3BC and ECC (F_2^m) Algorithm" Proceedings of the International Conference on e-Technology, e-Commerce and e-Service (EEE'04), IEEE 2004.

[10]. M. Peyravian, N. Zunic, "Methods for protecting password transmission", Elsevier Journal of Computers & Security, vol. 19, issue 5, pp. 466–469, July 2000.

[11] T. H. Chen, W. B. Lee, "A new method for using hash functions to solve remote user authentication", ACM Journal of Computers and Electrical Engineering, vol. 34, issue 1, pp. 53–62, Jan 2008.

[12]. Deng Jian-zhi, Cheng Xiao-hui, Gui Qiong, "Design of Hyper Elliptic Curve Digital Signature", IEEE International Conference on Information Technology and Computer Science, pp. 45-47, 2009.

[13]. S. Subasree and N. K. Sakthivel , " DESIGN OF A NEW SECURITY PROTOCOL", IJRRAS 2 (2), pp. 95-103, 2010.

[14]. Hung-Min Sun, and et al., "Dual RSA and its Security Analysis", IEEE Transaction on Information Theory, pp 2922 – 2933, Aug 2007.

[15]. Sida Lin and Qi Xie,"A secure and efficient mutual authentication protocol using hash function", International Conference on Communications and Mobile Computing, pp. 545-548, 2009

[16]. Sattar J Aboud, "Secure E-payment Protocol", International Journal of Security, (IJS), vol. 3, Issue 5, pp. 85-92, 2009.

[17]. Baddeley M, "Using e-cash in the new economy: An economic analysis of micro-payment systems", Journal of Electronic Commerce Research, vol. 5, no. 4, 2004.

[18]. J. Hubaux, and L Buttyan, "A micro-payment scheme encouraging collaboration in multi-hop cellular networks", In Proceeding of Financial Cryptography, LNCS 2742, Springer-Verlag, pp. 15–33, 2003.

[19]. N. Koblitz, "Elliptic Curve Cryptosystems," Math. Comp., vol. 48, pp. 203-209, 1987.

[20]. R. Rivest, "The MD5 message digest algorithm", Technical report RFC 1321, IETF, April 1992.

[21]. NIST, "Secure hash standard", Technical report FIPS 180-1, NIST, US Department of Commerce, April 1995.

[22]. T. ElGamal, "A public-key cryptosystem and a signature scheme based on discrete logarithms", IEEE Trans. Information Theory 31, pp. 469-472, 1985.

[23]. W. Diffie, M.E. Hellman, "New directions in cryptography", IEEE Trans. on Inform. Theory, vol. 22, pp. 644 - 654, 1976.

[24]. R.L. Rivest, A. Shamir, L. Adleman, "A method for obtaining digital signatures and public key cryptosystems", Commun. ACM vol. 21, pp. 120–126, 1978.

[25]. Fox C., "Essential Microsoft Operations Manage:" r. O'reily, 2006.

[26] N. Gura, A. Patel, A. Wander, H. Eberle, and S.C. Shantz, "Comparing Elliptic Curve Cryptography and RSA on 8-bit CPUs". Proceedings of Workshop on Cryptographic Hardware and Embedded Systems (CHES 2004), 6th International Workshop, pp. 119–132, 2004.

[27]. D. Bleichenbacher and A. May, "New attacks on RSA with small secret CRT-exponents," in Public Key Cryptology—PKC 2006, ser. Lecture Notes in Computer Science. New York: Springer, vol. 3958, pp. 1–13, 2006.

[28]. D. Boneh and G. Durfee, "Cryptanalysis of RSA with private key d less than N ," IEEE Trans. Inf. Theory, vol. 46, no. 4, pp. 1339–1349, Jul. 2000.

[29] E. Jochemsz and A. May, "A polynomial time attack on standard RSA with private CRT-exponents", 2007.

[30] M. J. Hinek, “Another look at small RSA exponents,” in Topics in Cryptology-CT-RSA 2006, ser. Lecture Notes in Computer Science, D. Pointcheval, Ed. New York: Springer, vol. 3860, pp. 82–98, 2006.

[31] X. Wang, Y.L. Yin, H. Yu, “Finding Collisions in the Full SHA-1”, Advances in Cryptology - CRYPTO 2005, Springer-Verlag,pp.17-36, 2005

[32]. XI Zhen-Yuan, CHEN He, WANG Xiang-Zhong, SHENG Jian-ling, FAN Yu-Tao, “Evaluation Model for Computer Network Information Security Based on Analytic Hierarchy Process”, Third International Symposium on Intelligent Information Technology Application, 2009.

[33]. Stallings, W. IEEE 802.11: Moving Closer to Practical Wireless LANs. IT Professional IEEE, Vol. 3, Issue 3, pp. 17 – 23, 2001.

[34]. Ayushi, “A Symmetric Key Cryptographic Algorithm”, International Journal of Computer Applications, Volume 1, No. 15, pp. 1- 4, 2010.

Author Biographies

First Author: Dr. Yogendra Kumar Jain presently working as head of the department, Computer Science & Engineering at Samrat Ashok Technological Institute Vidisha M.P India. The degree of B.E. (Hons) secured in E&I from SATI Vidisha in 1991, M.E. (Hons) in Digital Tech. & Instrumentation from SGSITS, DAVV Indore (M.P), India in 1999. The Ph. D. degree has been awarded from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2010. Research Interest includes Image Processing, Image compression, Network Security, Watermarking, Data Mining and has been published more than 40 Research papers in various Journals/Conferences, which include 12 research papers in International Journals.
Tel: +91-7592-250408,
E-mail: ykjain_p@yahoo.co.in.

Second Author: Nitin presently pursuing M. Tech. in CSE from Samrat Ashok Technological Institute Vidisha M.P India. He Secured M. C. A. from K.N.I.T. Sultanpur, UP India in 2005.
Tel: +91-9236114203,
E.mail: nitin4u.in.sati@gmail.com.
Research Interest includes Network security, Image Processing.

High Dimensional Data Classification through Attribute Reduction using RPCA Approach

Rajashree Dash[1], Rasmita Dash[2]

[1,2]Department of Computer Science and Engineering,
ITER, Siksha O Anusandhan University,
Bhubaneswar, Orissa, India
[1]rajashree_dash@yahoo.co.in, [2]rasmita02@yahoo.co.in

Abstract: Classification is one of the most commonly encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned to a predefined group or class based on a number of observed attributes related to that object. For many classification problems, a higher number of attributes used do not necessarily translate into higher classification accuracy. Hence attribute reduction can serve as a pre-processing tool of great importance before solving the classification problems. The main purpose is to reduce the maximum number of irrelevant features while maintaining acceptable classification accuracy. In this paper we proposed to use a hybridized RPCA approach of attribute reduction, which initially apply PCA to obtain reduced uncorrelated attributes specifying maximal variances in the data with minimum loss of information. Then we proposed to use Rough set theory on the PCA reduced data to discover discriminative features that will be the most adequate ones for classification. Lastly neural network has been applied for comparing the classification accuracy of some biological data sets with original attributes and reduced attributes.

Keywords: Data Classification, Feature Reduction, Feature Selection, Principal Component Analysis, Rough Set Theory.

1. Introduction

The ever increasing demand for a knowledge based system has focussed much of attention of researchers on knowledge acquisition. The task of extracting general knowledge from databases is known to be the most difficult part of creating a knowledge-based system. Data mining is a convenient way of knowledge extraction from large data sets and focusing on issues relating to their feasibility, usefulness, effectiveness and scalability. Classification is one of the most frequently encountered data mining tasks, used to assign an object into a predefined group or class based on a number of observed attributes related to that object. For many classification problems, a higher number of attributes used do not necessarily translate into higher classification accuracy. In some cases the performance of algorithms devoted to speed and predictive accuracy of the data characterization can even decrease. Therefore, attribute reduction can serve as a pre-processing tool of great importance before solving the classification problems. The main purpose is to reduce the maximum number of irrelevant features while maintaining acceptable classification accuracy.

Attribute reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality that corresponds to the intrinsic dimensionality of the data [1], [2]. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data. Dimensionality reduction approaches fall into two categories i.e. Feature Selection (FS) and Feature Reduction (FR). Feature Selection algorithm aims at finding out a subset of the most representative features according to some objective function in discrete space. Feature Extraction/ Feature Reduction algorithms aim to extract features by projecting the original high-dimensional data into a lower-dimensional space through algebraic transformations. It finds the optimal solution of a problem in a continuous space.

In this research, an approach of high dimensional data classification using neural network through attribute reduction using RPCA method has been proposed, in which initially PCA has been applied to obtain reduced uncorrelated attributes specifying maximal variances in the data and then the Rough set theory has been applied to generate the reduced set of necessary attributes or to construct the core of the attribute set by finding the upper and lower approximation of the reduced data set. This is a combination of feature selection approach with feature reduction to obtain a minimal set attributes retaining a suitably high accuracy in representing the original features. This approach will produce a reduced set of attributes which specify the maximal variances in the data as well as the discriminative features most adequate for classification, with minimum loss of information. Lastly the classification accuracy of some biological data sets with original and reduced attributes has compared using neural network.

2. Related Work

The problem of finding a reduced set of relevant features retaining a suitably high accuracy in representing the original features has been the subject of much research. Feature Selection using Rough sets theory is a way to identify relevant features, which has been validated by the improvement on the performance of the KNN classifier [3]. The classification accuracy of the classifiers intended for use in high dimensional domains can be increased by applying Principal Component Analysis which also increases its computational efficiency [4]. In [5] a new face recognition method based on PCA, LDA and neural network has been proposed specifying a high recognition rate. Rough set has also used for feature selection in medical data bases like Mammograms, HIV etc. without decision attribute with the application of clustering [6]. A novel method for dimensionality reduction of a feature set by choosing a subset of the original features that contains most of the essential information, using the same criteria as the ACO hybridized with Rough Set Theory has proposed in [8]. RST can only be applied on discretized data. A survey of discretization technique has been proposed in [13]. It has been validated that, the unsupervised methods like k-means clustering can perform equally well to that of supervised methods as it uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute.

3. Preliminaries

3.1 Principal Component Analysis

Principal Component Analysis [11],[12] is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation of the data that describes maximum variances in the data with minimum reconstruction error. It transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. Hence PCA is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set, without much loss of information.

PCs are calculated using the eigen value decomposition of a data covariance / correlation matrix or singular value decomposition matrix, usually after mean centering the data for each attribute. Covariance matrix is preferred when the variances of variables are very high compared to correlation. It would be better to choose the type correlation when the variables are of different types. Similarly the SVD method is used for numerical accuracy.

The transformation of the dataset to the new principal component axis produces the number of PCs equivalent to the no. of original variables. But for many datasets, the 1st several PCs explain the most of the variances, so the rest can be eliminated with minimal loss of information. The various criteria used to determine how many PCs should be retained for the interpretation are as follows:

- Using Scree Diagram plots the variances in percentage corresponding to the PCs which will automatically eliminate the PCs with very low variances.
- Fixing a threshold value of variance, so that PCs having variance more than the given threshold value will be retained rejecting others.
- Eliminate PCs whose eigen values are smaller than a fraction of the mean eigen value.

3.2 Rough Set Theory

Rough set theory is a new mathematical approach to imprecision, vagueness and uncertainty. It can be used for reduction of data sets, finding hidden data patterns and generation of decision rules. RST can be used as a tool to discover data dependencies and to reduce the no. of attributes contained in the data set using the data alone, requiring no additional information [9],[10]. Given a dataset with discretized attributes, it is possible to find a reduct of original attributes that are most predictive of the class attribute. Rough set reducts can be found by using degree of dependency or using discernibility matrix. In [7] a detailed concept of Rough set theory and decision tables for data analysis has provided.

4. Proposed Model

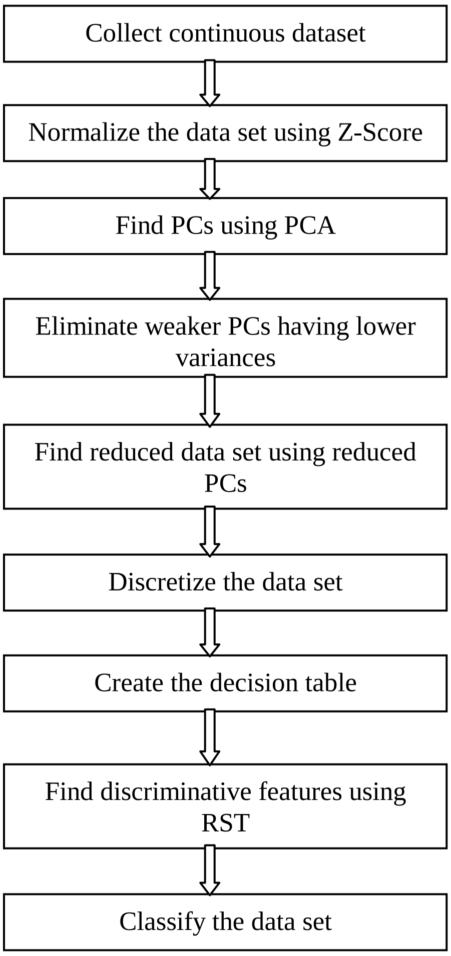

Collect continuous dataset
Normalize the data set using Z-Score
Find PCs using PCA
Eliminate weaker PCs having lower variances
Find reduced data set using reduced PCs
Discretize the data set
Create the decision table
Find discriminative features using RST
Classify the data set

Figure1. Data Classification through Attribute Reduction using RPCA Approach

Classification accuracy for high dimensional data may not be accurate most of the time due to noisy and outliers associated with original data. Also for some data the computational complexity increases rapidly as the dimension increases. Hence to improve the accuracy of classification, we proposed a method to apply PCA on original data set, so that the correlated variables exist in the original data set will be transformed to possibly uncorrelated variables, which are reduced in size, and then to apply the rough set theory on that reduced data set, which may contain some redundancy and to get the discriminative features. Before applying RST, we proposed to discretize the data set using a suitable unsupervised discretization technique. Lastly a suitable classification technique will be applied to the test data considering its reduced attributes to find its class value.

5. Experimental Analysis

The proposed method has been implemented on four biological data sets containing initially the continuous attributes i.e. Pima Indian Diabetes data set, Lung Cancer data set, Breast Cancer data set and SPECTF Heart data set, taken from the UCI machine learning repository using met lab. The data set details are given in the table 1.

Table 1. Data set Details

Data set	No. of Instances	No. of Attributes	No. of Class values
Pima Indian Diabetes	156	8	2
Breast Cancer	286	9	2
Spectf Heart	267	44	2
Lung Cancer	32	56	3

The experimental details are given in following steps.

Step 5.1 Data Normalization

Using the Normalization process, the initial data values are scaled so as to fall within a small specified range; so that any attribute having higher domain value will not dominate the attribute having lower domain value.

Step 5.2 Attribute reduction using RPCA

A set of PCs are calculated using Singular value decomposition of the normalized data. Then a transformation matrix is created containing the PCs having variances more than the mean variance, ignoring the other PCs and this transformation matrix is applied to the normalized data set to produce the new reduced projected dataset.

The reduced data set is discretized using an unsupervised discretization method. Here we preferred to use discretization using k-means clustering as it uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute and also it can perform equally well to that of supervised methods.

To apply RST, first a decision table containing object ids, the discretized attributes and the decision attribute is created. The class attribute of the data set has been considered as the decision attribute. Rough Set methods for finding reduct of attributes mainly categorized into two distinct approaches: those that incorporate the degree of dependency measure (or extensions), and those that apply heuristic methods to generate discernibility matrices. Although it is guaranteed to discover all minimal subsets using discernibility matrix method, it is a costly operation. Again simplifying discernibility function for reduct is a NP-hard problem. Hence here it is preferred to use the dependency based approach. Using Rough Set Theory the reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. Here we have first calculated the dependency of each attribute and then the best candidate has chosen. This process has continued till the dependency of the reduct equals the consistency of the data set. The reduced set of attributes obtained by applying RST on the discretized data set, with different discretization intervals has shown in the table 2.

Table 2. Reduced Attributes obtained through RPCA Approach

Data set	No. of original attributes	No. of reduced attributes obtained by PCA	No of reduced attributes obtained by RPCA with different discretization intervals			
			K=2	K=3	K=4	K=5
Pima Indian Diabetes	8	3	3	3	3	3
Breast Cancer	9	3	2	2	2	2
Spectf Heart	44	9	8	5	5	5
Lung Cancer	56	17	7	4	4	4

Step 5.3 Data Classification

The biological data sets have been classified using a feed forward back propagation network with two layers. One third of the data set has used as test data and the remaining as training data. Classifying the data set with original attributes, the PCA reduced attributes and with the reduced attributes obtained through RPCA approach, the classification accuracy obtained has shown in the table 3. In all cases the classification accuracy obtained by the reduced data set through RPCA approach is more than the original data set.

Table 3. Comparison of Classification Accuracy of Datasets

Data Set	Classification accuracy with original attributes.	Classification accuracy with PCA reduced attributes.	Classification accuracy with reduced attributes obtained by RPCA model .
Pima Indian Diabetes	58	64	64
Breast Cancer	76	81	86
Spectf Heart	85	71	85
Lung Cancer	65	68	74

6. Conclusion

In this paper some biological datasets with large no. of attributes have been classified through attribute reduction by RPCA approach. The attribute reduction through RPCA approach is a suitable combination of feature selection method with the feature reduction method. It has been implemented on the continuous data set, by applying an efficient PCA method, an efficient unsupervised discretization technique and a reduction algorithm of RST. As a result of which a no of uncorrelated and discriminative attributes, more adequate for classification has been obtained. These attributes also specifies the maximal variances among the dataset by retaining the original property of the data set. Again comparing the classification accuracy of the data sets using neural network, it was observed that the data classification through attribute reduction using RPCA approach provides more accuracy compared to the PCA reduced data and the original dataset, by retaining the original property of data set.

References

[1] Maaten L.J.P., Postma E.O., Herik H.J. van Den., "Dimensionality Reduction: A Comparative Review.," Tech. rep. University of Maastricht, 2007.

[2] Yan J., Zhang B., Liu N., Yan S., Cheng Q., Fan W., Yang Q., Xi W. and Chen Z., "Effective and Efficient Dimensionality Reduction for Large Scale and Streaming Data Preprocessing," IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 3, pp. 320-333, 2006.

[3] Frida Coaquira and Edgar Acuna, "Applications of Rough Sets Theory in Data Preprocessing for Knowledge Discovery," Proceedings of the World Congress on Engineering and Computer Science, pp.707-712, 2007.

[4] Changjing Shang and Qiang Shen, "Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study," International Journal of Computational Intelligence Research, Vol. 1, No. 1, pp. 68-76, 2005.

[5] Sahoolizadeh A. H., Heidari B. Z. and Dehghani C. H, " A New Face Recognition Method using PCA, LDA and Neural Network," World Academy of Science, Engineering and Technology, Vol. 41, pp. 7-12, 2008.

[6] Thangavel k. and Pethalakshmi A, "Feature Selection for Medical Database Using Rough System," AIML Journal, Vol. 6, No. 1, pp. 11-17, 2006.

[7] Nasiri J. H. and Mashinichi M, "Rough Set and Data Analysis in Decision Tables," Journal of Uncertain Systems, Vol. 3, No. 3, pp. 232-240, 2009.

[8] Mishra Debahuti, Rath Amiya Kumar and Acharya Milu, "Rough ACO: A Hybridized Model for Feature Selection in Gene Expression Data," International Journal of Computer Communication and Technology, Vol. 1, No. 1, pp. 85-98, 2009.

[9] Srivastava D. K, "Data Classification: A Rough - SVM Approach," Contemporary Engineering Sciences, Vol. 3, No. 2, pp. 77- 86, 2010.

[10] Lee-Chuan Lin, Zhu Jing, Junzo Watada, Tomoko Kashima and Hiroaki Ishii, " A Rough Set Approach to Classification and its Application for the Creative City Development," International Journal of Innovative Computing, Information and Control, Vol. 5, No. 12, pp. 4859-4866, 2009.

[11] Sampath Deegalla and Henrik Bostro, "Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods," Proceedings of the 8th International Conference on Intelligent data Engineering and Automated Learning, pp. 800-809, 2007.

[12] Valarmathie P., Srinath M. and Dinakaran K, "An Increased Performance of Clustering High Dimensional Data Through Dimensionality Reduction Technique," Journal of Theoretical and Applied Information Technolog*y*, Vol. 13, pp. 271-273, 2009.

[13] Sellappan Palaniappan and Tan Kim Hong, " Discretization of Continuous Valued Dimensions in OLAP Data Cubes," International Journal of Computer Science and Network Security, Vol. 8, No. 11, pp. 116-126, 2008.

Author Biographies

Rajashree Dash She is an assistant professor in computer science and engineering department of ITER, SOA University, Bhubaneswar, India. She has completed her MTECH from SOA University. Her area of research is data mining, software engineering.

Rasmita Dash She is an assistant professor in computer science and engineering department of ITER, SOA University, Bhubaneswar, India. She has completed her MTECH from SOA University. Her area of research is data mining, software engineering.

Steganographic Tools for BMP Image Format

Prof.Sumedha Sirsikar Prof. Asavari Deshpande

Department Of Information Technology
MAER'S Maharashtra Institute of Technology
Pune, Mahatashtra, 411038, India.
{sirsikarsd, asavari.deshpande}@gmail.com

***Abstract*:** The goal of steganography is to transmit a message through some innocuous carrier i.e. text, image, audio and video over a communication channel where the existence of the message is concealed. In this paper we present characteristics, performance, and robustness of various Steganographic freeware tools. Out of this few tools are used for steganography and steganalysis that evaluate and identify the shortcomings which are useful to Forensic analysts. Performance measurement is carried out on the basis of visual inspection and statistical comparison. The result for few tools is presented in this paper.

Keywords: Steganography, BMP format, Tools

1. Introduction

Private and personal data communication is a need of today's world. A solution to this is a Cryptography [1] that scrambles the confidential information which can be read by only intended recipient. But the communication is easily recognized because of encrypted data and hence privacy and confidentiality is lost. Alternate technique of hiding information is Steganography that provides privacy and security by making confidential communication invisible. Cryptography is not related to the invisible communication, where the goal is to secure communication from an eavesdropper. But steganography hides the existence of the communication channel itself. The output of steganography operation is not apparently visible whereas in cryptography output is scrambled, hence it can draw attention.

In literature the term 'information hiding' is often used as a synonym for steganography. Steganography need to be robust against distortion like compressions or color adjustment. Also, it communicates in a completely undetectable manner unlike watermarking.

It is used in the field of secret communication ex, exchange of highly confidential data in a covert manner say on public discussion board or forum. It can also be used for secure and invisible storage of confidential information like patents that can be stored on hard disk partitions.

The objective of this paper is to analyze and carry out statistical study of various Steganography tools (freeware). To support this experimental results are carried out which identifies their characteristics.

The rest of the paper is organized as follows. Section 2 introduces history technical aspect of steganography. Section 3 gives perspective of Steganography. Section 4 details of various tools for data hiding. Section 5 describes comparative results of tools. Section 6 conclusion and future work.

2. History of Steganography

Steganography is derived by Johannes Trithemus (1462-1516) from "Steganographia" and comes from the Greek word, defined as "covered writing". Hidden message will not arouse an eavesdropper's suspicion.

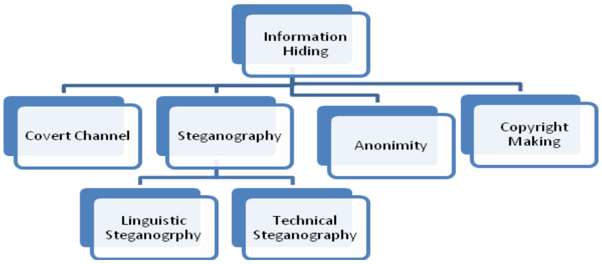


Figure 1: Classification of Information Hiding Techniques

As shown in Figure 1 information hiding has many subparts, one of the most important is steganography. The Linguistic steganography is defined by Chapman et al. [25] as "the art of using written natural language to conceal secret messages". Here the cover medium is composed of natural language text and the text itself which can be generated to have a cohesive linguistic structure. In the Technical steganography, carrier is physical medium such as microdots and invisible inks rather than a text.

2.1 Technical aspect of information hiding

There are three aspects in information hiding systems which contend with each other: capacity, security and robustness [17]. Capacity refers to the amount of information that is hidden in the medium. However, security is important when a secret communication is kept to be secret and undetectable by eavesdroppers. Robustness can be explained as the amount of modification the stego-medium can withstand before an adversary can destroy the hidden information.

Information hiding techniques are used, to fuse the digital content within the image regardless of the different file formats and the status of the image (digital or analog). For example, extra data about an image is added in special tags such as file headers. This information will be lost when the image is printed. As headers are tied to the image as long as image exists in digital form. Hence secret data is not communicated using file headers.

Various techniques are used to implement Steganography, which are used in different tools such as Least Significant Bit (LSB), manipulation of image and compression algorithms, and modification of image properties such as luminance [2].

3. Perspective of Steganography

Internet publishes images to convey ideas for mass communications that acts as a good carrier for hiding information using various tools. These tools are divided into two groups: Image domain ex LSB and Transform domain.

In LSB technique, change in 1 or 2 bit is unnoticeable to the human eye . LSB has a limitation on amount of secret message to be added into cover image. If it exceeds stego-image would appear to be suspicious .Tools used in this group are StegoDos, S-Tools, Mandelsteg, EzStego, Hide and Seek, Hide4PGP, Jpeg-Jsteg, White Noise Storm, and Steganose[1][3][4][5]. Here lossless image formats are used where data is directly manipulated and recovered for ex. Windows Bitmap (BMP). Due to the use of reliable compression algorithm hidden message is not lost.

The transform domain group that involves manipulation of algorithm and image transform for ex. Discrete Cosine Transformation (DCT) and Wavelet transformation. For hiding data it uses more significant area of cover images and luminance of image. Examples of tools are PictureMarc, JK-PGS, SysCop, SureSign. These techniques are more robust than bit-wise techniques. JPEG uses DCT to get image compression for ex. Jpeg-Jsteg tool.

Both image and transform domain characteristics are present in few techniques for example pattern block encoding, spread spectrum methods and masking [2].The addition of redundancy to the hidden information protects data from image processing methods such as cropping and rotating. In this paper we are giving details of various tools belongs to image domain.

4. Evaluation of Steganalysis tools

4.1 S-Tool 4.0

S-Tool [6] reduces the total number of 256 colors to 32 colors. Then basic colors are expanded over several palette entries sorted by their luminance. Though the block of colors appears to be same but it differs by 1-bit value. Same approach is used for color and gray scale images. The stego-image produced from gray scale image no longer remains a gray scale image as the RGB value within pixel may vary by 1-bit. It works with 24-bit images. It can hide 115,184 bytes of data. Encryption algorithm like IDEA, DES, Tripledes, MDC are used for encryption of secret data. Steganography algorithm implemented here uses concept of color reduction (average color, average pixels) as large RGB and luminosity distance. It also enables floyed Steinberg dithering concept.

4.2 Steghide 0.5.1

StegHide [7] hides data into JPEG and BMP. It has features like compression of embedded data, encryption of embedded data and automatic integrity checking using a checksum. The default encryption algorithm used is Rijndael with a key size of 128 bits (AES) in the cipher block chaining mode. The checksum is calculated using the CRC32 algorithm.

Color and sample frequencies are not changed. Graph theoretic approach is used to implement steganography algorithm. Secret data is first compressed and then encrypted. By using pseudo random number generator positions of the pixels from cover image is determined to embed secret data.

Those pixels that need to be changed are sorted out. Then graph theoretic matching algorithm is used to find pair of positions where embedding is carried out. The pixels at the remaining positions (not part of pairs) are also modified by overwriting pixels to contain embedded data. Exchanging pixel values imply that the first order statistics (number of times color occurs in the picture) is not changed.

4.3 Hide In Picture (HIP):

HIP [8] hides any kind of file inside standard bitmap pictures by modifying its color in a way that is almost unnoticeable by the human eye. For too large hidden files noise is inserted in the stego image. One color of the picture may be set as a transparent color; nothing is stored in that area. It also supports for overwriting hidden data. Blowfish (default) and Rijndael encryption algorithm and checksum is also supported. For 24-bit images the size of hidden file should not more than 40% of the picture size.

4.4. wbStego4.3

wbStego4 [9] is used for Windows95/98/ME, and Windows NT 4.0/2000/XP. In it information about the carrier file (e.g. copyright - information) is added without using a separate file. Any data, texts, graphics or even executable programs can be hidden in the carrier files by slight changed, so that the manipulation is not detected. This version of wbStego4 uses bitmaps (*.BMP), text files (*.TXT), HTML files (*.HTM) and Adobe™ Portable Document Format (*.PDF) as carrier files. It also offers cryptographic methods.

Steganographic algorithm uses the concept of color depth (how many bit per pixel define the color value). For higher security it requires more color. Very large areas in one color Bitmap should be avoided. The color depths and amount of data hidden are shown in Table 1.

Table 1: Number of color and amount of hidden data

Bitmap Size	**Number of colors or gray scales**	**Amount of hidden data (size of bitmap in byte : size of hidden data)**
4 bit	16	4 : 1
8 bit	256	8 : 1
24 bit	16,777,216	8 : 1

It also uses lossless compression (e.g. PCX) technique, but the color depth may not be changed. Lossy JPEG may not be used. To reduce the size of the carrier file, which is very important for online transmission, to reduce the size of carrier file any compression utility (e.g. ARZ, LZH, PKZIP, WinZip) can be used, as they are lossless.

4.5 CryptaPix 3.05

CryptaPix ™ [10] is an image file management and encryption program for Windows. In CryptaPix steganographic functions are built around a pixel shuffling routine. AES random number generator is used. The plaintext data is divided into 3-bit segments. Those are overwritten into the lowest red, green, and blue bits in the selected pixels. A maximum of 3 bits (out of the 8 available) for each color channel may be used.

4.6 StegoStick 1.0

StegoStick [11] hides any kind of file like BMP, GIF, and JPEG and the result is in the BMP format. Perceptible distortion is not observed in the stego-image. It uses encryption techniques like DES, TripleDES, and RSA.

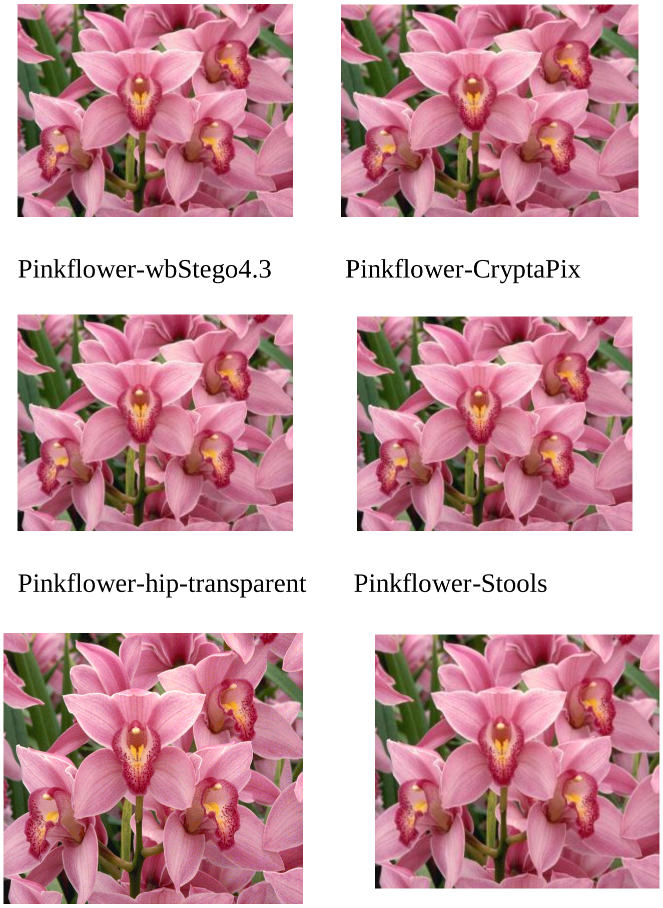
Table 2: Features of Steganographic tools

Name of Tool	**Cover image**	**Utility type**	**Encryption Techniques**	**Steganlysis Algorithm**
S-Tool 4.0	GIF BMP	GUI	IDEA,DES, Triple DES, MDC	1-bit LSB & color reduction
Steghide 0.5.1	JPEG BMP	Commandline	DES,TDES, Blowfish, Rijandael,RC2 , All block ciphers, CRC32	Graph theoretic approach using pixel pair exchange
Hide in picture (HIP) 2.1	BMP	GUI and Commandline	Blowfish, Rijandael, CRC	one colour : transparent (Index 88 RED 0) that does not hide secret data
wbSteg o4.3	BMP	GUI	DES,TDES, RSA	Color depth
CryptaPix 3.05	BMP	GUI	AES	Secure data division is into 3-bit segment. Bit/pixel/color
StegoStick 1.0	JPEG , BMP, GIF	GUI	DES , Triple DES , RSA	---

All the Steganographic tools studied in this paper are summarized in the Table 2. Stego-images for all tools are generated in BMP formats except for Steghide (BMP, JPEG) tool. All tools support Windows but StegoStick, StegHide, and HideinPix supports linux.

4. Comparative results of Tools

Test and experiments are carried out using all above tools, with the help of cover image of size 301 kb, and secret data (file hidden.txt) of size 9 bytes. That results into all stego-images which is shown in the Table 3. Here we have observed the size of the cover image and stego-image is slightly varied. All stego-images are shown in the Figure 2.



Pinkflower-wbStego4.3 Pinkflower-CryptaPix

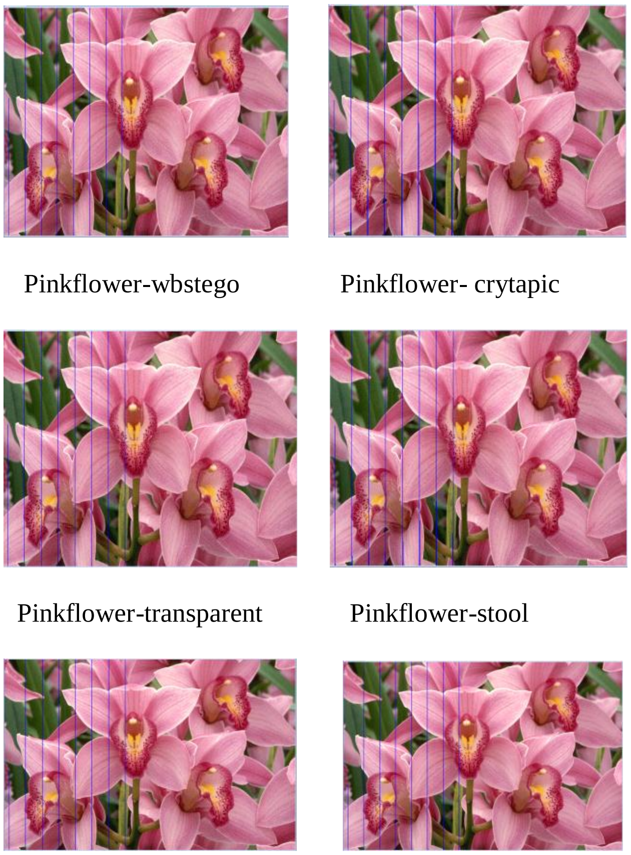
Pinkflower-hip-transparent Pinkflower-Stools

Pinkflower-StegoHide Pink flower StegoStick

Figure 2. Stego-images resulted from various tools

Peak-Signal-to-Noise-Ratio (PSNR) is used as a major of performance for image distortion. PSNR is expressed o a logarithmic scale in decibels (dB). PSNR values below 30dB indicates that the distortion caused by embedding secret data is very low.

Data distribution of quantitative variables is displaced graphically using technique called as Histogram as shown in Figure 3. In case of images variables are nothing but intensity values of image. Histogram is used to test the presence of any abnormalities observed in the stego-image as compared to the cover image.



Pinkflower-wbstego Pinkflower- crytapic

Pinkflower-transparent Pinkflower-stool

pinkflower_stegoHide pinkflower_stegostick

Figure 3. Histogram of Stego Images

Table 3. Comparative study of Steganographic tools

Steganographic Tools	**Image Size in (Kb)**	**Name of stego-file**	**Size of Stego-file**	**PSNR**
StegoStick 1.0	301KB	pinkflower_Tdes_stego	1.17MB	17.05961257
S-Tool 4.0	301KB	pinkflower_stool_stego	900KB	17.03729143
StegHide0.5.1	301KB	pinkflower-stego-StegHide	900KB	17.07062074
Hide In Picture (HIP) 2.1	301KB	pinkflower-stego-HIP	301KB	17.03717922
WbStego4.3	301KB	pinkflower-stego-wbstego	301KB	17.03719218
CryptaPix3.05	301KB	pinkflower-stego-cryptapix	900KB	17.03736634

6. Conclusion and Future work

Information hiding techniques has become important in security research. Invention of applications and technologies brings new threats that require new protection mechanisms. Hence for ex applications such as e-banking, e-trading, mobile telephony, medical data interchanging etc., requires study of Steganograpic tools. Thus this paper throws light on the various features of these tools.

This study can be further extended for JPEG, GIF, PNG file formats. The result of this paper can guide the steganalyst to develop a tool which can automatically extract hidden messages in images.

References

[1] S. Das, Subhendu Das, B Bandyopadhyay, and S Sanyal, "Steganography and Steganalysis: Different Approaches". Internatioanl Journals of Computer Applications,Volue No7, Number 9,year 2010.

[2] Neil F. Johnson, and Sushil Jajodia, "Steganalysis of Images Created Using Current Steganography" Software, Springer Verlag (1998), proceedings for Second Information Hiding Workshop in Portland, Oregon, USA, April 15-17, 1998.

[3] Ahmed Ibrahim, "Steganalysis in Cmputer Forensics", Security Research Centre Conferences "Australian Digital Forensics Conference",available at http://ro.ecu.au/adf/10, Dec Year 2007.

[4] Abbas Cheddad, Joan Condell, Kevin Curran, and Paul McKevitt, "A Comparative Analysis of Steganographic

Tools”, Information technology and Telecommunication Conference 2007 ,pp 29-36

[5] Pedram Hayati, Vidyasagar Potdar, and Elizabeth Chang, “A Survey of Steganographic and Steganalytic Tools for the Digital Forensic Investigator”, Workshop of Information Hiding and Watermarking with IFIPTN New Brunswick, Canada, July 2007.

[6] S-Tool4.0: ftp://ftp.funet.fi/pub/crypt/mirrors/idea.sec.dsi.unimi.it/code/s-tool4.zip

[7] Steghide0.5.1.: http://steghide.sourceforge.net/

[8] Hide in Picture http://sourceforge.net/projects/hide-in-picture/

[9] wbStego4.3: http://members.xoom.com/wbailer/wbstego/index.htm

[10] CryptaPix 3.05 http://www.briggsoft.com

[11] StegoStick1.0 http://www.infosectechnologies.com/StegToolandWatermarkingTable.pdf

Author Biographies

Sumedha D Sirsikar sumedha.sirsikar@mitpune.edu.in has pursued M.E. Computer Engineering from College of Engineering, Pune, Maharashtra, India in year 200. She is presently working as a Professor and Head of Information Technology department in MIT Pune. Her area of interest is computer network and information security.

Asavari A. Deshpande (asavari.deshpande@gmail.com) has pursued B.E in Computer Science and Engineering from M.G.M. college of engineering , Nanded, Maharashtra, India in year 2004.She is presently perusing Masters in Computer engineering from 2008.and working as lecturer in MAEER’s M.I.T Pune Her area of interest is database normalization.

www.ijcset.excelingtech.co.uk