January - June, 2013 Vol. 5 No. 1 ISSN 0973-5658

BIJIT

BVICAM's International Journal of Information Technology

Indexed with EBSCO (USA), Google Scholar (USA), DOAJ (Sweden) & Open J-Gate (USA)

CONTENTS

Bharati Vidyapeeth's
Institute of Computer Applications and Management
A-4, Paschim Vihar, Rohtak Road, New Delhi-63
Email : bijit@bvicam.ac.in, Website : http://www.bvicam.ac.in

Our Indexing at International Level

EBSCOhost Electronic Journals Service (EJS) is a gateway to thousands of e-journals containing millions of articles from hundreds of different publishers, all at one web site. For further details, click at http://www.ebscohost.com/titleLists/tnh-coverage.htm

Open J-Gate is an electronic gateway to global journal literature in open access domain. Launched in 2006, Open J-Gate is aimed to promote OAI. For further details, click at http://informindia.co.in/education/J-Gate-Engineering/JET-List.pdf

DOAJ aims at increasing the visibility and ease of use of open access scientific and scholarly journals, thereby promoting their increased usage and impact. For further details, click at http://www.doaj.org/doaj?func=issues&jId=87529&uiLanguage=en

Google Scholar provides a simple way to broadly search for scholarly literature and repositories from across different parts of the world. For further details, click at http://scholar.google.com/scholar?hl=en&q=BIJIT%2BBVICAM&btnG=

Stanford Libraries provide world' best scholarly published resources for academic and research purposes. For further details, click at http://searchworks.stanford.edu/?utf8=%E2%9C%93&q=BVICAM&search_field=search&utf8=%E2%9C%93

Cabell's Directory of Publishing Opportunities contains a wealth of information designed to help researchers and academics, match their manuscripts with the scholarly journals which are most likely to publish those manuscripts. For further details, click at https://ssl.cabells.com/index.aspx

Academic Journals Database is a universal index of periodical literature covering basic research from all fields of knowledge. For further details, click at http://journaldatabase.org/journal/issn0973-5658

The Open Science Directory is a search tool for open access journals mainly for developing countries. For further details, click at https://atoz.ebsco.com/titles/searchresults/8623?GetResourcesBy=TitleNameSearch&Find=BVICAM&SearchType=Contains

Indian Citation Index (ICI) is an abstracts and citation database, with multidisciplinary objective information/knowledge contents from about 1000 top Indian scholarly journals For further details, click at. http://www.indiancitationindex.com/htms/release_notes.htm

The Florida Institute of Technology (Florida Tech or FIT), is a Research University located in Melbourne, Florida, (USA). Its library resources are considered as one of the best research facility in the world. For further details, click at https://catalog.lib.fit.edu/Record/3117339/Details

WorldCat is the world's largest network of library content and services. It has a repository of over 1.5 billion items aimed to help researchers, academicians and practitioners. For further details, click at http://www.worldcat.org/search?q=BIJIT&qt=results_page

and many more..., for more details click at http://www.bvicam.ac.in/BIJIT/indexing.asp

Volume 5, Number 1

January – June, 2013

BIJIT - BVICAM's International Journal of Information Technology is a half yearly publication of Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), A-4, Paschim Vihar, Rohtak Road, New Delhi – 110063 (INDIA).

ISSN 0973 – 5658

Address for Correspondence:

Prof. M. N. Hoda
Editor-in-Chief, BIJIT
Director, Bharati Vidyapeeth's
Institute of Computer Applications and Management (BVICAM),
A-4, Paschim Vihar, Rohtak Road, New Delhi – 110063 (INDIA).
Tel.: 91 – 11 – 25275055 Fax: 91 – 11 – 25255056 E-Mail: bijit@bvicam.ac.in
Visit us at www.bvicam.ac.in/bijit

Published and printed by Prof. M. N. Hoda, Editor-in-Chief, BIJIT and Director, Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), A-4, Paschim Vihar, New Delhi – 63 (INDIA). Tel.: 91 – 11 – 25275055, Fax: 91 – 11 – 25255056.
E-Mail: bijit@bvicam.ac.in; mca@bvicam.ac.in Visit us at www.bvicam.ac.in/bijit

Editorial Board

Editorial

It is a matter of both honor and pleasure for us to put forth the ninth issue of BIJIT; the BVICAM's International Journal of Information Technology. It presents a compilation of eleven papers that span a broad variety of research topics in various emerging areas of Information Technology and Computer Science. Some application oriented papers, having novelty in application, have also been included in this issue, hoping that usage of these would further enrich the knowledge base and facilitate the overall economic growth. This issue shows our commitment in realizing our vision "to achieve a standard comparable to the best in the field and finally become a symbol of quality".

As a matter of policy of the Journal, all the manuscripts received and considered for the Journal by the editorial board are double blind peer reviewed independently by at-least two referees. Our panel of expert referees posses a sound academic background and have a rich publication record in various prestigious journals representing Universities, Research Laboratories and other institutions of repute, which, we intend to further augment from time to time. Finalizing the constitution of the panel of referees, for double blind peer review(s) of the considered manuscripts, was a painstaking process, but it helped us to ensure that the best of the considered manuscripts are showcased and that too after undergoing multiple cycles of review, as required.

The eleven papers that were finally published were chosen out of seventy nine papers that we received from all over the world for this issue. We understand that the confirmation of final acceptance, to the authors / contributors, sometime is delayed, but we also hope that you concur with us in the fact that quality review is a time taking process and is further delayed if the reviewers are senior researchers in their respective fields and hence, are hard pressed for time.

We further take pride in informing our authors, contributors, subscribers and reviewers that the journal has been indexed with some of the world's leading indexing / bibliographic agencies like ***EBSCO*** *(USA),* ***Open J-Gate*** *(USA),* ***DOAJ*** *(Sweden),* ***Google Scholar, WorldCat*** *(USA),* ***Cabell's Directory*** *of Computer Science and Business Information System (USA),* ***Academic Journals Database, Open Science Directory, Indian Citation Index, etc.*** *and listed in the libraries of the world's leading Universities like* ***Stanford University, Florida Institute of Technology, University of South Australia, University of Zurich****, etc. Related links are available at http://www.bvicam.ac.in/bijit/indexing.asp. It will certainly further increase the citations of the papers published in this journal thereby enhancing the overall research impact.*

We wish to express our sincere gratitude to our panel of experts in steering the considered manuscripts through multiple cycles of review and bringing out the best from the contributing authors. We thank our esteemed authors for having shown confidence in BIJIT and considering it a platform to showcase and share their original research work. We would also wish to thank the authors whose papers were not published in this issue of the Journal, probably because of the minor shortcomings. However, we would like to encourage them to actively contribute for the forthcoming issues.

The undertaken Quality Assurance Process involved a series of well defined activities that, we hope, went a long way in ensuring the quality of the publication. Still, there is always a scope for improvement, and so, we request the contributors and readers to kindly mail us their criticism, suggestions and feedback at bijit@bvicam.ac.in and help us in further enhancing the quality of forthcoming issues.

Editors

CONTENTS

Online Rule Generation Software Process Model

Rajni Jain[1], Satma M C[2], Alka Aroa[3], Sudeep Marwaha[4] and R C Goyal[5]

Submitted in April, 2012; Accepted in October, 2012

Abstract - For production systems like expert systems, a rule generation software can facilitate the faster deployment. The software process model for rule generation using decision tree classifier refers to the various steps required to be executed for the development of a web based software model for decision rule generation. The Royce's final waterfall model has been used in this paper to explain the software development process. The paper presents the specific output of various steps of modified waterfall model for decision rules generation.

Index Terms - Software Process model, Modified waterfall model, Decision Rule, Decision Tree

1. INTRODUCTION

Classification is the discovery of a predictive learning model that classifies a data item into one of several predefined classes [4]. The classification model can predict the class of objects whose class label is unknown. It is also called classifier [8,16]. Patil et. al. have done work on fault classification of mechanical System using self organizing techniques [16]. Verma et. al. have used rough set techniques for 24 hour knowledge factory [17]. But none of these algorithms are available online. We have made attempt to develop a software process model for online rule generation. The model can be used by other researchers for their own algorithms to make online software.

A decision tree is a classifier expressed as a recursive partition of the instance space. It consists of nodes that form a rooted tree. The leaf nodes denote class labels or class distribution. The non-leaf nodes denote a test on an attribute and branches denote outcome of the test [12]. An online rule generation software is required by researchers and data mining personnel who have to generate rules to facilitate the development of expert systems or pattern recognizing in various domains. Presently researchers use their individual program running on their desktop [18]. Online software enables the easy access to it using the default browser on the client machine. But it is not available yet. So there is a need to develop the online decision rule generation software referred to as 'GenRule'.

To build software, it is important to go through a series of predictable steps. The steps are like a roadmap that helps to develop a high quality system. This roadmap is also called a software process.

[1]*NCAP, Pusa, New Delhi-12*
[2, 3, 4, 5] *IASRI, Pusa, New Delhi-12*
E-mail: [1]rajni@ncap.res.in and [2]satmaktm@gmail.com

The software development life cycle (SDLC) is the entire process of formal, logical steps taken to develop a software product [13]. There are five phases that are part of the SDLC [3]. These phases are requirements definition, design, coding, testing and maintenance. SDLC models are created based on the order in which they occur and the interaction between them [5]. The modified waterfall model developed by Royce has been used in the development of GenRule.

The present paper is an attempt to identify and document the requirements for developing the online software described above. The rest of the paper is organized as follows. The section 2 presents the software process model concepts. Its sub-sections deal with the various phases of SDLC namely requirement analysis, design, coding, testing and maintenance. Section 3 presents the conclusion.

2. SOFTWARE PROCESS MODEL

A software process model is an abstract representation of the architecture, design or definition of the software process [14]. There are varieties of software development process models to show how organizing the process activities can make the development more effective [1]. One of the basic software process models is waterfall model. But it is not flexible. Its phases are strictly linear [9]. So the Royce's modified final waterfall model has been used in the development of the rule generation software using decision tree classifier.

The advantage of the modified waterfall model is that it is a more relaxed approach to formal procedures, documents and reviews. It also reduces the huge bundle of documents. Due to this, more time can be devoted to coding without bothering about the procedures. This in turn helps to finish the product faster [9]. The different phases in the Royce's final waterfall model [15] with reference to the development of GenRule are explained in the subsequent sub sections.

2.1 Requirement Analysis

Requirements are set of functionalities and constraints that end-user expects from the system. For GenRule, users are mainly developers of expert systems, students and data mining researchers who are interested in generating rules from data. Recently, expert systems are being developed for various agricultural crops like wheat, maize, mustard etc. In production systems like expert systems, knowledge is required to be fed in the form of rules. Usually these rules are made at the expense of valuable time of experts. In the field of agriculture, vast amounts of research data are generated every day and to convert those huge amounts of data to useful and knowledgeable decision rules that can help in crucial decision making, decision rule generation software is required. Consequently the experts could spend their time only on

validating the generated rules that are provided along with accuracy measures. Rule generation using decision tree classifier is providing additional benefits of visualization effect in the form of decision tree, which adds to the understandability of rules.

Information about user's requirements was gathered from literature [11,7] and also by consulting the prospective users like researchers involved in development of expert systems. It helped to know what should be accomplished by the application. These requirements were further analyzed for their validity and the feasibility of incorporating them in GenRule were explored.

User requirements are broadly categorized into two types namely functional and non-functional requirements. Functional requirements describe what the software should do. They include user requirements, input requirements, computational requirements, output requirements, exception handling etc. Non-functional requirements refer to the requirements that are not directly concerned with the specific functions delivered by the system [14]. They are broadly categorized into performance requirements and system requirements. The first one deals with the level of performance required by users. It includes various other requirements like usability, human factor and security issues etc. The system requirements for GenRule are identified at three levels namely, client level, server level and at the programmer level.

2.1.1 Functional Requirements

The sequence diagram facilitates the understanding of user requirements [14]. On the basis of interaction with various categories of users, sequence diagram for GenRule is presented in Fig 1. It explains the sequence of actions to be followed by user to generate rules using GenRule. The sequence diagram clearly exhibits the following functional requirements of the users.

i. The user has to be validated by checking the user name and password. Only the valid users should have the facility to access the software.

ii. Input Requirements: Facility should be provided to input the data in excel or CSV (Comma Separated Values) file format. The user has to enter the partition preference of the dataset and select the required attributes from the available attributes and finally the target attribute from the selected attributes.

iii. Computational Requirements:

(a)The input data should be validated for non-categorical and missing values. If they are present, exception will occur and error message will be displayed. There should be partition of input data into training and test dataset randomly according to the partition preferences given by the user. It should have the provision to select the required attributes from the whole set of attributes. It should also provide facility to the user to select the class attribute from the already selected list of attributes.

(b)User should have the facility to classify future data instances if its classifier model is already built in the software. The user may be allowed to choose a classifier already built and stored by GenRule. If the model is not learned, user has to generate the rules first and go for prediction of future instances. (c)Data flow model is an intuitive way of showing how data is processed by a system. The data flow diagram shown in Fig.2&3 illustrates how the data flows through a sequence of processing steps in GenRule to generate decision rules. The graphical user interface is well explained with the help of use-case diagram (Fig. 4-7). Use-case diagram identifies the user interactions with the software. The various interactions involved are described in Fig.5, 6 and 7 respectively. The generation of rules from the training data using ID3 algorithm in the process 1.5 (Fig.2) is further explained using the data flow diagram in Fig 3.

iv.Output Requirements: Facility should be provided to display the generated rules and the corresponding decision tree view along with evaluation measures like rule coverage, rule accuracy, precision, recall, F-measure, confusion matrix, training accuracy and test accuracy. Exporting and saving facility should be provided in Excel, text and XML file format for the generated rules for further implementation. For improving the understandability of the rules, the corresponding decision tree should be displayed and there should be provision to save it in XML format.

v.On logging out, user should return to the home page.

2.1.2 Non-Functional Requirements

i. The software should be friendly and available on the internet, with the authentication of user name and password.

ii. The software should meet all kinds of user requirements efficiently.

iii. It should provide accurate output in all aspects.

iv. Online help facility should be included.

v. Results should be reliable.

vi. Client level specification: Any browser with latest facility like IE6 or higher, Excel 2003 or 2007.

vii. Server level specification: Windows 7, IIS 7.0, Microsoft .NET Framework Version 3.0, 2 GB RAM, 2.53 GHz Processor, 320 GB Hard Disk,

viii.Programmer level specification: Microsoft Windows 7, Visual Studio 2008, IIS 7.0, 2 GB RAM, Excel 2007, IE8, 320 GB Hard Disk.

The ID3 algorithm is a recursive algorithm for building decision tree and decision rules [10]. As a part of requirement analysis, it is important to get an understanding of the basic ID3 algorithm used. The ID3 algorithm is explained using flow chart shown in Fig. 8.

2.2 Design

The requirement specifications from first phase were studied in this phase and system design was prepared. System design helped in specifying the hardware and system requirements and also helped in defining the overall system architecture. A set of software design concepts has evolved over the history of software engineering [14].

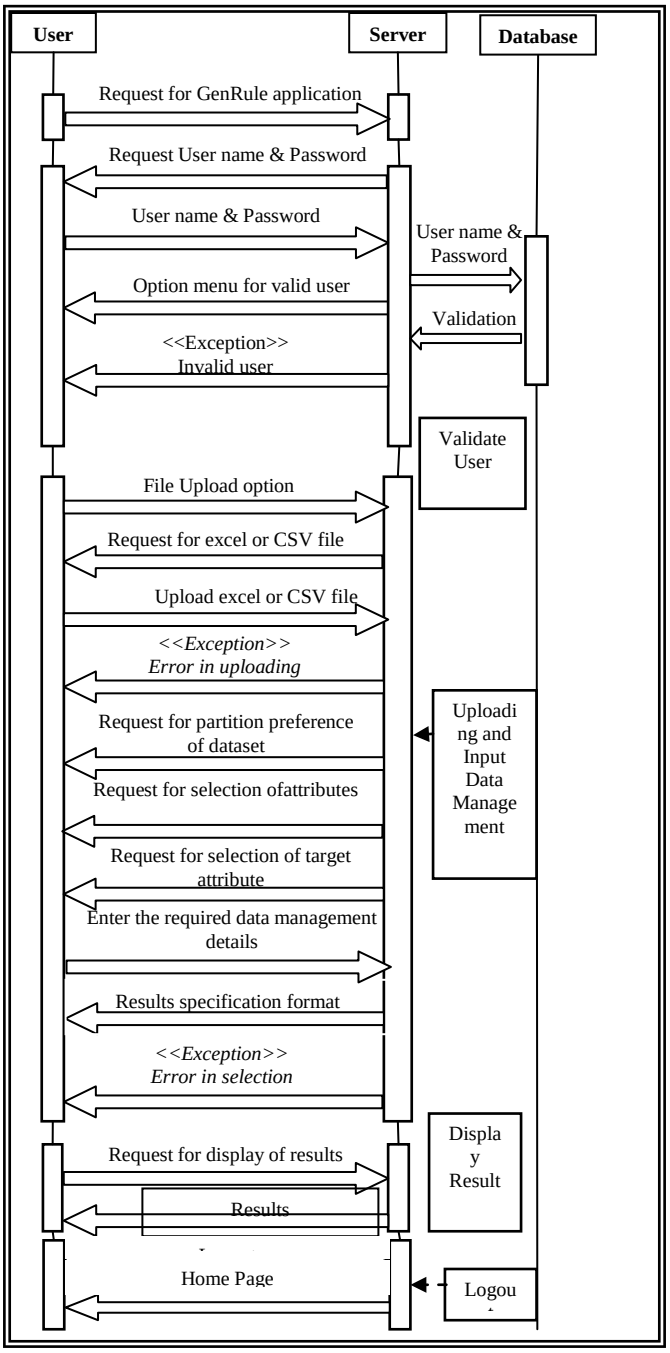


Figure 1: Sequence diagram for rule generation

The design of GenRule is presented with the help of input design, output design, database design and design of modules. Modularity is one of the powerful concepts for software design. Most complex design tasks are solved by breaking them down into manageable part called modules [2].

2.2.1 Input Data Design

Data may be entered to the software using Excel or CSV file. In the data, the columns should represent the attributes and the rows should contain the dataset instances. One of the attributes should be the class attribute. Table1 represent one such sample dataset where lodging is class attribute.

Variety	Crop_duration	Climate	N_fertilizer	Lodging
resistant	early	rainy	low	no
resistant	early	rainy	high	no
tolerant	medium	rainy	low	yes
susceptible	early	rainy	low	yes

Table 1: Sample dataset

Database Design

Database for the system is maintained using MS SQL server. Database should contain a table that is useful to store the generated classifier for subsequent use (Table 2).

Column Name	DataType	Allow Nulls
id	int	no
Parent_id	int	yes
Node name	varchar(50)	yes

Table 2: Classifier table schema

The database constructed corresponding to the sample input data given in the Table 1 is shown in Table 3. The last column denotes the name of the nodes. The id for each node is given in the first column. The parent_id column gives the parent node of each node. If parent_id is zero, it is a root node. This classification table is useful to classify the unseen cases.

id	Parent_id	Node name
1	0	variety
2	1	resistant
3	2	climate
4	3	dry
5	4	yes
6	3	rainy
7	6	no
8	1	susceptible
9	8	yes
10	1	tolerant
11	10	N_fertilizer
12	11	high
13	12	no

Table 3: Classifier table for crop lodging dataset

There is a database namely 'aspnetdb' that contains tables with the login details of users under the sql membership provider functionality of ASP.NET.

2.2.2 Output Design

Outputs for GenRule software are decision rules in table format and decision tree in tree view format. It also computes various evaluation measures like confusion matrix, precision, recall, F-measure, training accuracy and test accuracy. The decision rules output table contains rule id and corresponding to each rule id, the class attribute column and other attributes column with respective values for the rule. The value of '*' given to attributes represents any value of the given attribute. Each rule is associated with its evaluation measures coverage and

accuracy, as coverage explains the data instances covered by a rule and accuracy explains the validity of a rule in the data instances. The coverage and accuracy of ith rule ($Rule_i$) can be computed using the given formula.

$$Coverage(Rule_i) = \frac{ncovers(Rule_i)}{|training\ dataset|}$$

$$Accuracy(Rule_i) = \frac{ncorrect(Rule_i)}{ncovers(Rule_i)}$$

Where ncovers ($Rule_i$) represents the number of data instances satisfying the antecedent of $Rule_i$ and ncorrect ($Rule_i$) represents the number of data instances correctly classified by $Rule_i$.The decision rule output table of the data given in Table 1 should be like the Table 4. Rules should be displayed in classic if-then format also (Table 5). A confusion matrix contains information about actual and predicted classification done by a classification system [6]. The performance of such systems is commonly evaluated using the data in the confusion matrix. Confusion matrix can be worked out for both training and test dataset. All the performance measures computed are functions of the confusion matrix (Table 6). Precision, recall and F-measure can be computed for the two class values.

Id	lodging	variety	Crop_duration	climate	N_fertilizer	Rule Coverage (%)	Rule Accuracy (%)
1	yes	susceptible	*	*	*	28	100
2	yes	resistant	*	dry	*	14	100
3	no	resistant	*	rainy	*	21	100
4	no	tolerant	*	*	high	14	100
5	yes	tolerant	*	*	low	21	100

Table 4: Decision rule table for crop lodging dataset

Rule_id	Decision Rule	Rule Coverage (%)	Rule Accuracy (%)
1	If [variety='susceptible'], then lodging='yes'	28	100
2	If [variety='resistant'], [climate='dry'], then lodging='yes'	14	100
3	If [variety='resistant'], [climate='rainy'], then lodging='no'	21	100
4	If [variety='tolerant'], [N_fertilizer='high'], then lodging='no'	14	100
5	If [variety='tolerant'], [N_fertilizer='low'], then lodging='yes'	21	100

Table 5: classic if-then rule for crop lodging dataset

Class values	Recall	Precision	F-measure	Accuracy (%)
yes	0.5	1	0.66	66
no	1	0.5	0.66	

Table 6: performance measures on crop lodging classifier

2.3 Coding

The functionalities of GenRule can be achieved by developing various modules and assigning different tasks (Table 7).
The implementation of the defined modules in the design phase was done using classes and methods (Table 8).

2.4 Testing, Integration and Maintenance

Each of the modules (Table 7) should be tested for their functionality individually during the unit testing for the desired output. These units should be integrated into a complete system during integration phase and should be tested to check if all modules/ units coordinate between each other and the system as a whole behaves as per the specifications. Bottom up integration was used for combining various modules [2].
The software has to be maintained as long as it is used for various applications. There should be proper documentation to facilitate the operation and maintenance as and when required.

Module Name	Description
Registration	Provide facility of sign up to new user
Login	Provide facility of login to users
Update	An option for change of password
Decision Rule Generation	The main module, which generate decision rules from the training dataset along with rule evaluation measures like rule coverage and accuracy
Decision Rule Validation	The module which validates the generated decision rules using the test dataset and provide evaluation measures
Decision Tree	Constructs the decision tree corresponding to the generated rule set

Module Name	Description
Classification	Provide the prediction of target attribute values for the future datasets which are unclassified. It is useful if the classifier is preexisting.
Prediction	Provide the prediction of target attribute values for the future datasets which are unclassified, successively after building the classifier
Sample Data	Provide sample data for the user
Contact Us	Provide contact details of the development team
Help	Provide online help about software

Table 7: Description of various modules in GenRule

Class name	Description
getSheetName	Establish connection to Excel file and get all sheet names from Excel file
fillInDataSet	Data present in various sheets of Excel file is uploaded into different dataset
DStoDT	Convert dataset to corresponding data table
Attribute	Designate column names as attributes with certain properties
ID3	Class that performs the ID3 algorithm functionality
TreeNode	Create tree node for allocating the attributes coming out of the ID3 algorithm

Table 8: Description of different classes in GenRule

3. CONCLUSIONS AND FUTURE SCOPE

Software process model explains the requirement specifications, input design, output design, database design, implementation, testing and maintenance phases. In this paper the software development life cycle of decision rule generation software is presented from its conception to its maintenance and implementation stage. The process model helped to develop GenRule software that can fulfill the requirements of agriculturalresearchers, teachers, students and other data mining personnels handling huge amounts of data. The model will also be helpful in future for any enhancements or maintenance of the software.

The software process explained above can be utilized for many other decision tree algorithms.

REFERENCES

[1]. S. T. Acuna, A. D. Antonio, X. Ferre, M. Lopez, and L. Mate. *The Software Process: Modelling, Evaluation and Improvement, Handbook of Software Engineering and Knowledge Engineering*. World Scientific Publishing Company.2000.

[2]. A. Goswami, N. Arora, and A. Sharma. *Fundamentals of Software Engineering*. Lakhanoal Publishers, Amritsar, India.2010.

[3]. N. Jenkins. *A Software Testing Primer*. Creative Commons, California.2008.

[4]. M. Kantardzic**.** *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley and Sons, Inc Publication, New Jersy.2011.

[5]. R. Kissel, K. Stine, M. Scholl, H. Rossman, J. Fahlsing, and J. Gulick. *Information Security*. NIST Special Publication, Gaithersburg. 2008.

[6]. R. Kohavi, and F. Provost. Glossary of Terms. *Machine Learning*, Vol. 30, No.3: 106-110.1998.

[7]. P. Langley and H. A. Simon. *Applications of Machine Learning and Rule Induction*. Institute for the Study of Learning and Expertise.1995.

[8]. A. M. Mahmood, N. Satuluri, and M. R. Kuppa. An Overview of Recent and Traditional Decision Tree Classifiers in Machine Learning. *Indian Journal of Research and Reviews in Ad Hoc Networks*, Vol.1,No.1: 10-12. 2011.

[9]. N. M. A. Munassar, and A. Govardhan. A Comparison Between Five Models of Software Engineering. *IJCSI International Journal of Computer Science*, Vol.7, No.5: 94-101.2010.

[10]. J. R. Quinlan. Induction of Decision Trees. *Machine learning*, Vol.1, No.1: 81-106.1986.

[11]. J. R. Quinlan. *Generating Production Rules from Decision Trees*. Massachusetts Institute of Technology, USA.2007.

[12]. L. Rokach, and O. Z. Maimon. *Data Mining with Decision Trees: Theory And Applications*. World Scientific Publishing Co Pvt Ltd. 2008.

[13]. W. Scacchi. *Process Models in Software Engineering*. John Wiley & Sons, Inc, New York. 2001.

[14]. I. Sommerville. *Software Engineering: A Practitioner's Approach*. Tata McGraw Hill. 2009.

[15]. J. Xiong. *New Software Engineering Paradigm Based on Complexity Science*. Springer Publication. 2011.

[16]. J. K. Patil, P.B. Ghewari and S.S. Nagtilak. Iterative Self Organised Data Algorithm for Fault Classification of Mechanical System, BIJIT - BVICAM's International Journal of Information Technology, issue 5, 3(1). 2011.

[17]. N. Verma, N. Nerma and A.B. Patki. Rough Set Technique for 24 Hour Knowledge Factory. BIJIT - BVICAM's International Journal of Information Technology, issue 7, 4(1), 2012.

[18]. N. Aggarwal, N. Prakash, S. Sofat. Mining Techniques for Integrated Multimedia Repositories: A Review, BIJIT - BVICAM's International Journal of Information Technology, Issue 1, 1(1), 2008.

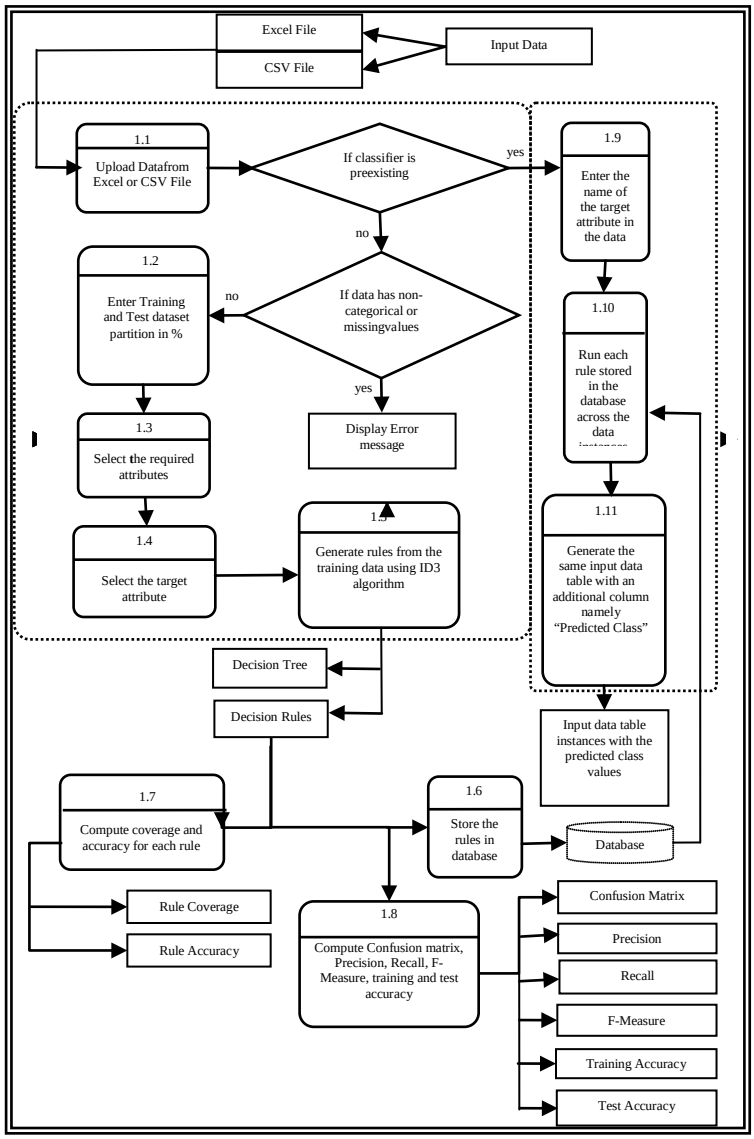


Figure 2: Data flow diagram for GenRule

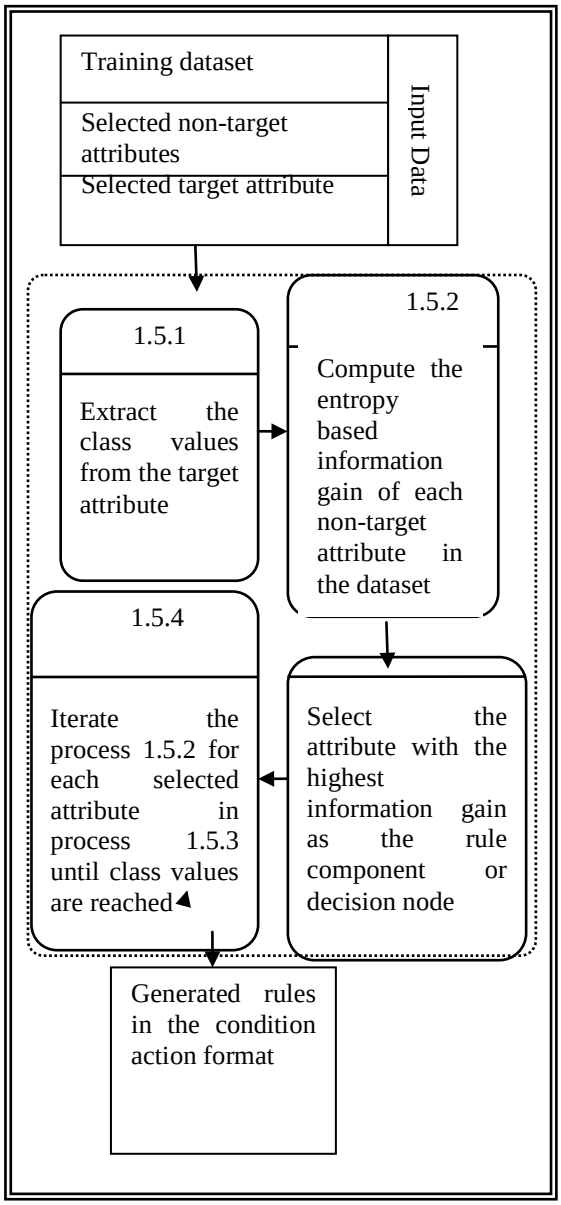


Figure 3: Data flow diagram for process 1.5 in Figure 2

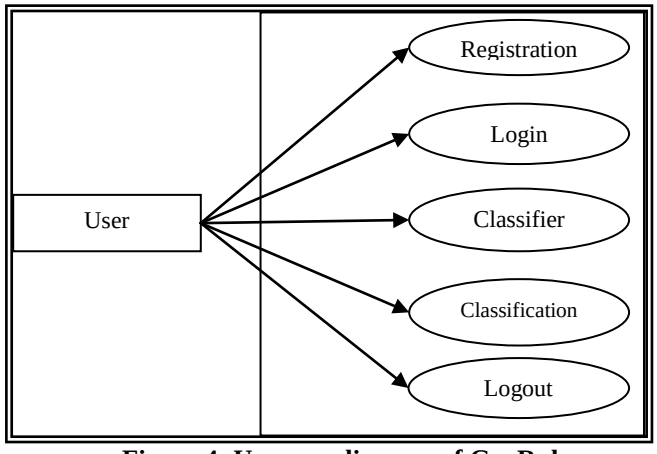


Figure 4: Use-case diagram of GenRule

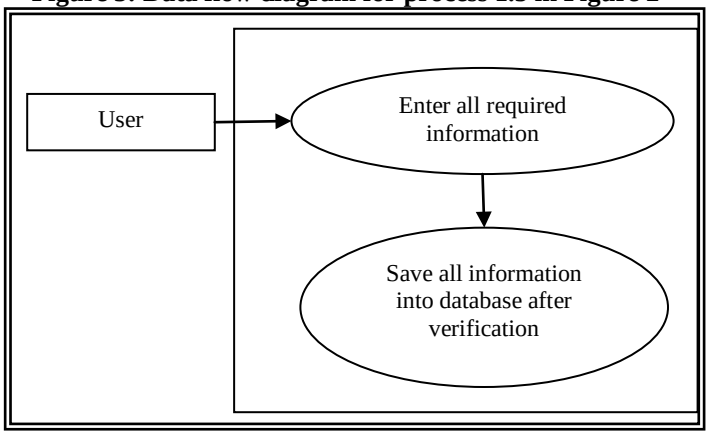


Figure 5: Use-case diagram for registration

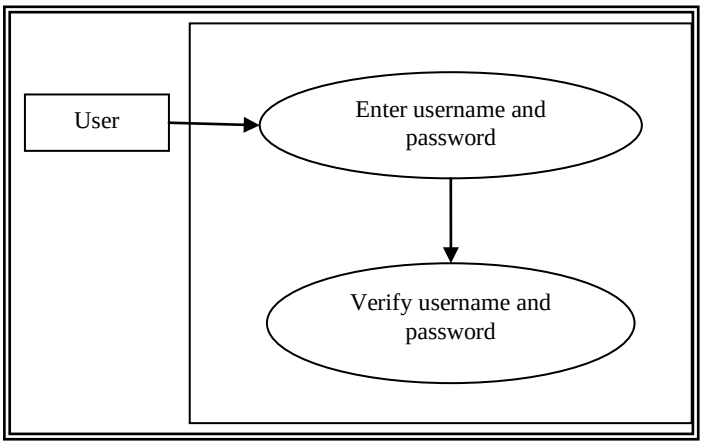


Figure 6: Use- case diagram for login

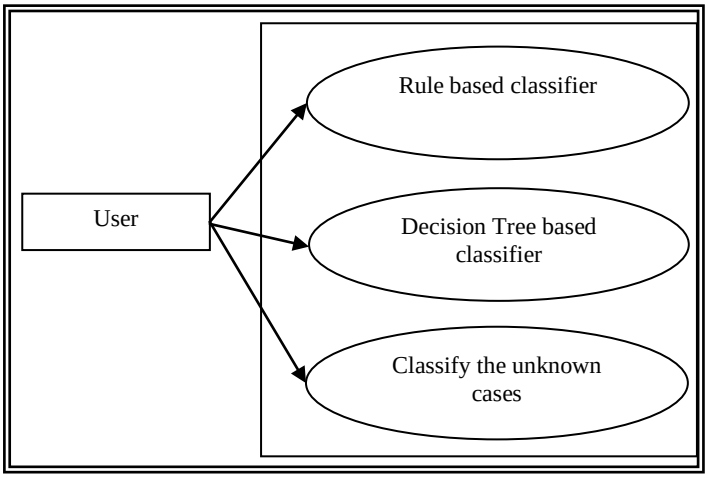


Figure 7: Use-case diagram for classification

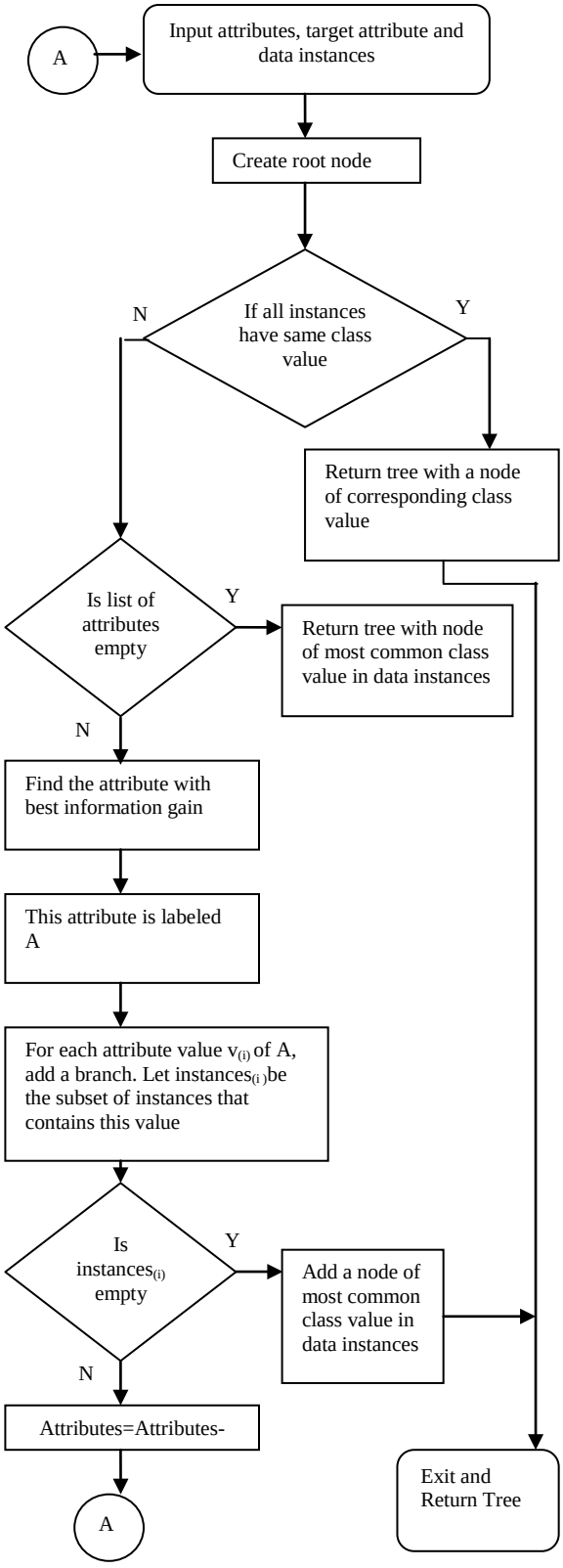


Figure 8: ID3 Algorithm Flowchart

Design and Implementation of Monophones and Triphones-Based Speech Recognition Systems for Voice Activated Telephony

Rupayan Das[1] and **Pradip K. Das**[2]

Submitted in April 2012; Accepted in January 2013

Abstract - Speech recognition is the ability of a machine or program to convert spoken words into its equivalent text form. Nowadays, most recognition systems use Hidden Markov Models for modeling the spoken utterances. In this paper we have implemented two speaker independent speech recognition systems which include all the words required for dialing a phone. The systems contain 42 words including digits from zero to nine and also include names of 20 persons. A total of 16,800 utterances have been used for training each system. The two systems are able to recognize continuous speech and it is implemented with the help of monophones and triphones using HTK. Experimental results show an accuracy of 74.11% for monophones based models and 93.77% for triphones based models.

Index Terms - HMM, HTK, Monophones, Triphones, Mel Frequency Cepstral Coefficient (MFCC).

1. INTRODUCTION

Pattern recognition is an important area of machine learning domain. The domain of pattern recognition is itself quite wide and encompasses several other interesting areas. The basic goal of a pattern recognition problem is to be enable a machine to identify as to which class, among a set of given classes, does a test pattern belongs. One interesting application of this area is presented in [1] for generation of traffic models in urban areas. [2] presents an interesting research on the problem of face recognition, which is now-a-days widely used as a measure to authenticate the users. [3] presents a very nice review of the statistical pattern recognition methods.

A subset of the pattern recognition domain is the area of speech recognition where spoken utterances are the patterns that are intended to be recognized. The process of speech recognition involves the communication between persons and machines where automata is generated to report the written equivalent of spoken words. From 1950's researchers were trying to make a device that can recognize human voice. In 1952, at Bell Laboratories a system for isolated digits recognition was built by Davis Biddulph and Balashek. The system heavily relied on the spectral resonances of the vowels of each digit. After that, lot of work on speech recognition has been done all over the world.

[1, 2] *Department of Computer Science & Engineering, Indian Institute of Technology (IIT), Guwahati, Assam, India*
E-mail:[1] *rupayan@iitg.ernet.in and* [2]*pkdas@iitg.ernet.in*

Some speech recognition systems give very good accuracy of more than 95% and are able to transcribe more than 150-160 words per minute. The improvement in the speech recognition system is increasing rapidly day by day. Nowadays, many hand-held devices like mobile phones, iPods, iPhones are trying to provide a good recognition system and research is still going on to improve the quality of the recognition accuracy. In the present era, mainly Hidden Markov Models (HMM) based speech recognition systems are used. HMM is a doubly embedded stochastic process with an underlying stochastic process that is not directly observable but can be observed only through another set of stochastic processes that produce the sequence of observations [4]. HMMs were first discussed in the second half of 1960 in a series of statistical papers by Leonard E. Baum and his colleagues [5]. In 1970 it has been first used as a tool for speech recognition by Baker [6] at CMU and by Jelinek and his colleagues at IBM [7]. Since then, due to its strong mathematical structure it gained its popularity day by day and started to be used in a wide range of applications, such as handwriting recognition [8], natural language domain and also for forecasting stock prices for interrelated markets [9], etc. HMM can also be used for speech recognition in other languages. In 2006 Gupta made an isolated word speech recognition for Hindi digits using continuous HMM [10]. Also in 2011 Kumar and Aggarwal made a Hindi recognition system using HTK which recognized 30 Hindi words [11]. In 2011 Hguyen presented a paper which describes a study of building a Vietnamese speech recognition system using HTK. The system gives the accuracy of 71.37% for speaker independent recognition before speaker adaptation and 75.96% after speaker adaptation [12]. HTK has also been used for speech recognition for other international languages such as Arabic language [13].

In this paper, a speaker independent recognition system is implemented with the help of Hidden Markov Model Toolkit which can be used to recognize continuous speech. The system includes all the commands required for dialing a phone. It consists of numbers from zero to nine and commands like "Call", "Dial", "Phone", "Flash", "Hangup", "Hash", "Star", "Redial" and "Hold". It also contains names of 20 persons. A total of 42 words have been used to make the system. The system is implemented using both monophones and triphones as base units. Experimental results show that the accuracy based on triphones models is much higher than the monophones based system.

2. HIDDEN MARKOV MODEL TOOLKIT (HTK)

HTK is a software toolkit for building and manipulating systems that use continuous density Hidden Markov models

(HMMs) [14]. It is a collection of library modules written in C combining which a system can be designed. The first version of the HTK was developed at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED) in 1989 by Steve Young. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. The software supports HMMs using both continuous density mixture Gaussians and discrete distributions and can be usedto build complex HMM systems [15].

2.1 HTK Implementation Structure

The different steps for building the HMMs using the toolkit are detailed below:

•Data Preparation: In this phase a database has been created by collecting data from 20 different speakers. Each speaker has 20 utterances of each word having a total of 16800 (20*42*20) utterances. The data is recorded using CSL workstation in a laboratory environment. A distance of approximately 5-10 cm is used between mouth of the speaker and microphone. Sounds are recorded at a sampling rate of 16000 Hz. After recording has been done, all the words are manually labeled and stored with a logical name.

•Feature extraction: As it is very complex to work with raw speech data, it is important to extract all relevant acoustic information in a compact form from raw speech. We use Mel frequency cepstral coefficients (MFCCs) [16] to extract feature vectors from the recorded raw data

•Model Training: In this phase, the first thing required is to define a prototype model which contains the information about the characteristics and the topology of the HMM. For our system, the topology used is 3-state left–right with no skips [17]. With the help of this proto file we generate the first HMM and then repeatedly re-estimate it to get the required optimal model.

3. IMPLEMENTATION DETAILS

First of all we make a grammar file which describes the words to be recognized. The grammar file for the telephone based system contains:

$digit = ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINE | ZERO;

$name = RAM | JOHN | AMIT | HEMANT | BIKASH | GOPAL | ARUN | SUMIT | JAMES | NITIN | MAYANK | DEBANJAN | ROHIT | ANIL | RAJA | STEVE | JHONSON | KRISHNA | NIL | PUNIT ;

$mode = ON | OFF;

(SENT-START (DIAL $digit $digit $digit $digit $digit $digit $digit $digit $digit $digit | (PHONE | CALL) $name | SPEAKER $mode | FLASH | HANGUP | HASH | STAR | REDIAL | HOLD) SENT-END)

From this grammar file, some sample commands that can be formed are listed in Table 1:

A total of 42 words have been selected to make the recognition system. The word CALL and PHONE can be used interchangeably. It is taken to give the user more flexibility while calling. After making the grammar it is saved in the *gram*file. The symbol $ denotes a string variable, the vertical bars denotes alternatives and the angle braces denotes one or more representations. After making the grammar file we need to make the word network for these words. This is done by executing the command **Hparse gram wdnet** which will take *gram* file as input and generate the word network file *wdnet* that contains each word-to-word transition.

Command	**Command**
DIAL 9985345631 (any 10 digit from zero to nine)	HASH
CALL RAM (any name chosen from $name in the grammar file)	STAR
PHONE RAM (any name chosen from $name in the grammar file)	REDIAL
FLASH	HOLD
HANGUP	

Table 1: Sample commands using the grammar file

The next step is to build the list of phonemes for each of the words in the vocabulary. This is done by using the command HDMan -m -w wlist -n monophones1 -l dlog dict names which will take as input names and wlist and generate the list of phonemes in the file monophones1.The wlist file contains the list of words and names file is same as wlist except that it also contains the phoneme sequences of the words. Table 6 contains all the words along with their corresponding phonemes. Now the silence sil is added to the list and saved in file monophone0. In addition SENT-END and SENT-START is augmented.

In order to train the system with the given words, the list of words to be spoken for training is generated. It is generated by using the command HSgen -l -n k wdnetdict > trainprompts which will use wdnet and dict files and generate the train prompts that contain a total of k training sentences. Next, the recording of all these sentences are to be done using the software HSLab provided by the toolkit.

Now for training the system it is required to replace the word in train.mlf file with its corresponding phonemes. This is done by executing the command HLEd -l '*' -d dict -i phones0.mlf mkphones0.led train.mlf which will take as input train.mlf file, dict file and mkphones0.led file and generate the corresponding phonemes in the file phones0.mlf. The train.mlf file contains the trainprompts sentences and mkphones0.led file contains commands used to replace the word with its corresponding phonemes.

The next step is to parameterize the raw speech waveforms into sequences of feature vectors. This is done by the command HCopy -T 1 -C cfg_mfc -S code_mfc.scp. The command will take code_mfc.scp and cfg_mfc file as input. The scp file contains the location of the .wav files and also the location of the .mfc files to be created. The configuration file cfg_mfc can be set as shown in Table 2 below:

Parameters	**Value**
TARGETKIND	MFCC_0_D_A
TARGETRATE	100000.0

Parameters	Value
SAVECOMPRESSED	T
SAVEWITHCRC	T
WINDOWSIZE	250000.0
USEHAMMING	T
PREEMCOEF	0.97
NUMCHANS	26
CEPLIFTER	22
NUMCEPS	12
ENORMALISE	F

Table 2: Contents of the *cfg_mfc* file

Now the monophones HMM is generated by using following steps:

For training the HMM, first of all a proto file is defined which defines the model topology. In our experiments the topology used is a 3-state left-right with no skips. The command HCompV -C config -f 0.01 -m -S tr_mfc_mono.scp -M hmm0 proto is executed to generate a new version of file proto and Vfloor in hmm0 directory. The config file contains the only line TARGETKIND = MFCC_0_D_A and tr_mfc_mono.scp file contain the locations of the MFC files. After that the model formed which is saved in the file proto is placed against each phoneme entry in hmmdefs file. Also copy the contents of vfloors to a file named macro.

Then it is required to re-estimate the flat start monophones models and this can be done by executing the command HERest -C config -I phones0.mlf -t 250.0 150.0 1000.0 -S tr_mfc_mono.scp -H hmm0\macros -H hmm0\hmmdefs -M hmm1 monophones0 which will generate hmmdefs and macros files in hmm1 directory. Executing the command two more times, the file hmmdefs and macros can be generated in hmm3 directory. The previous step generates a 3 state left-to-right HMM for each phone and also a HMM for the silence model sil. Now we need to create a 1 state short pause sp model by copying the contents of the sil model and placing it in the sp model. Since sp has its emitting state tied to the center state of the silence model, the centre step is retained and other states are deleted.

Now for making the model more robust, it is required to add an extra transition in the sil model which absorbs the various impulsive noises in the training data. This can be done by executing the command HHEd -H hmm4\macros -H hmm4\hmmdefs -M hmm5 sil.hed monophones1 where the input files are monophones1 and sil.hed. The sil.hed file contains data including:

```
AT 2 4 0.2 {sil.transP}
   AT 4 2 0.2 {sil.transP}
   AT 1 3 0.3 {sp.transP}
  TI silst
   {sil.state[3],sp.state[2]}
```

The AT command adds transitions to the given transition matrices and TI command creates a tied-state called slist. When we execute the command HHED, we get corresponding hmmdefs and macros files in the hmm5 directory. Finally, another two passes of HEREST are applied using the phone transcriptions with sp models between words. This results the models to be stored in hmm7 directory.

Since the dictionary contains multiple pronunciations of some words, so the phone models created so far can be used to realign the training data and create new transcriptions. This can be done by executing the command HVite -l '*' -o SWT -b silence -C config -a -H hmm7/macros -H hmm7/hmmdefs -i aligned.mlf -m -t 250.0 -y lab -I train.mlf -S train.scp dict monophones1 which uses the HMMs stored in hmm7 to transform the input word level transcription train.mlf to the new phone level transcription aligned.mlf using the pronunciations stored in the dictionary dict. When the aligned.mlf file is created, we execute another two passes of HERest which will store the required HMMs in hmm9 directory.

Now we are ready to run the recognizer for live input. For this, a configuration file config2 is needed which will convert the input data into its parameterization form. The config2 file contains the following parameters and their values:

Parameters	Value
SOURCERATE	625.0
SOURCEKIND	HAUDIO
SOURCEFORMAT	HTK
TARGETKIND	MFCC_0_D_A
TARGETRATE	100000.0
ENORMALISE	F
USESILDET	T
MEASURESIL	F
OUTSILWARN	T

Table 3: Contents of the config2 file

Now for recognizing the word, the command HVite -H hmm9/macros -H hmm9/hmmdefs -C config2 -w wdnet -p 0.0 -s 5.0 dict monophones1 is used which uses a token passing algorithm to perform viterbi-based speech recognition.The Viterbi algorithm finds the best state sequence for the observation sequence obtained from the previous steps. It takes wdnet, dict, monophones1and a set of HMMS as input. It converts the word network to a phone network and then attach the appropriate HMM definition to each phone instance. When we run the command, it first measures the speech and background silence level by prompting the user to speak an arbitrary sentence. After that it will repeatedly recognize the word and output into the terminal.

The triphones based HMMs are generated with the help of the following additional steps:

First the command HLEd -n triphones1 -l '*' -i wintri.ml fmktri.led aligned.mlf is executed which will convert the monophone transcriptions in aligned.mlf to an equivalent set of triphone transcriptions in wintri.mlf. Also a list of triphones is saved in triphones1 file. The mktri.led file is an edit script.

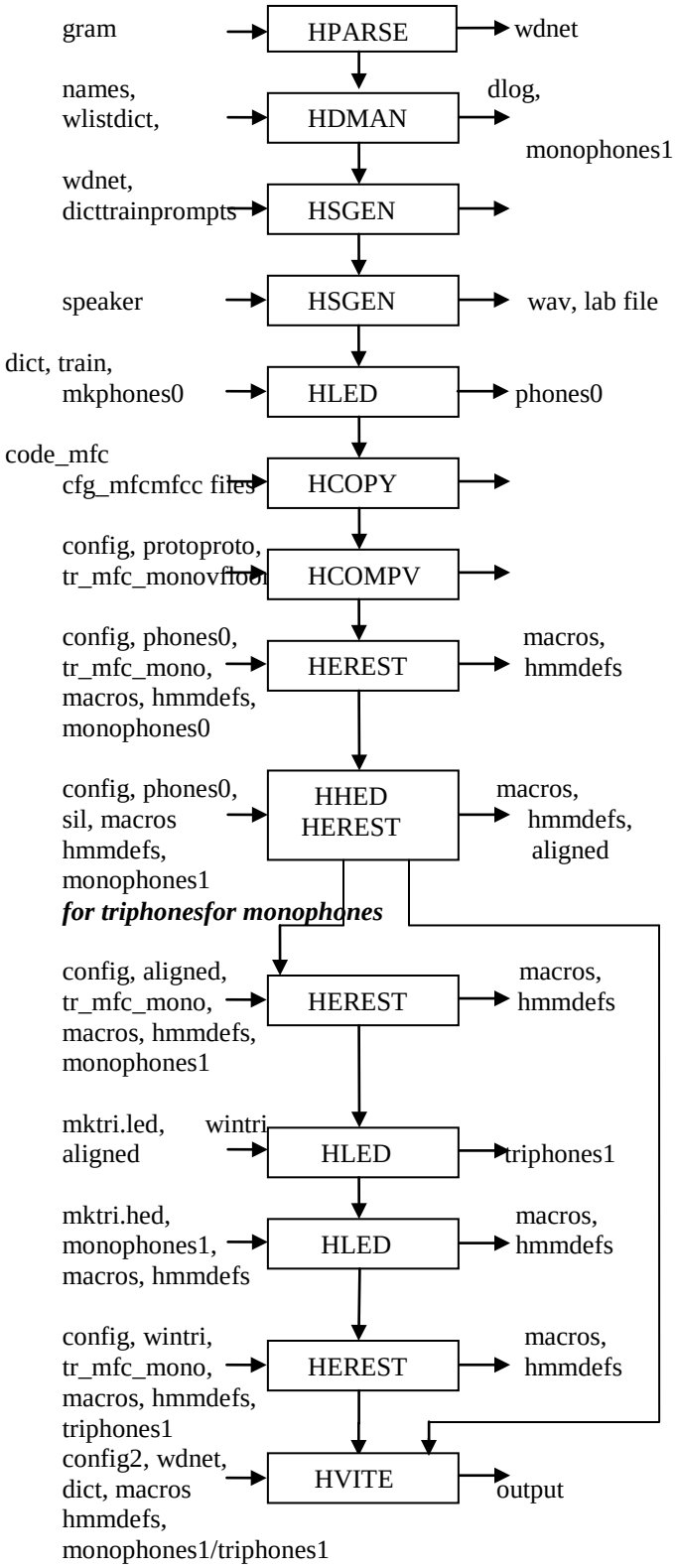


Figure 1: Block diagram representing the working of monophones and triphones-based recognition system.

which contains WB sp, WB sil and TC where WB commands define sp and sil as word boundary symbols. Now we have to make an edit script mktri.hed containing a clone command CL followed by TI commands to tie all the transition matrices in each triphone set. Now the cloning of models can be done by using the command HHEd -B -H hmm9/macros -H hmm9/hmmdefs -M hmm10 mktri.hed monophones1. Finally, another three passes of command HERest -B -C config -I wintri.mlf -t 250.0 150.0 1000.0 -s stats -S train.scp -H hmm11/macros -H hmm11/hmmdefs -M hmm12 triphones1 are applied to save the resultant models in hmm13 directory.
For live recognition, we can use the command:

HVite -H hmm13/macros -H hmm13/hmmdefs -C config2 -w wdnet -p 0.0 -s 5.0 dict triphones1

The complete training and recognition process is explained with the help of a block diagram in Figure 3.

4. EXPERIMENTAL RESULTS

To find the accuracy of the system, 20 speakers have been selected to test the system. From these 20 speakers, 10 speakers are those whose voices have been already included while making the model and 10 new speakers are included to test the system. Each person speaks 20 utterances of each word resulting in a total of 400 (20*20) utterances per person. The system is tested for both monophones and triphones based models and the results are shown in Table 4 below:

Sl. No.	Words	Recognition accuracy (in percentage)	
		Monophones based HMMs	Triphones based HMMs
1	AMIT	77.25	94.25
2	ANIL	74.5	95.5
3	ARUN	71.5	92
4	BIKASH	76.25	93
5	CALL	80.25	97
6	DEBANJAN	77.75	95.5
7	DIAL	81.75	96.25
8	EIGHT	67.75	91.75
9	FIVE	73.75	93.5
10	FLASH	64.5	92.25
11	FOUR	78	93
12	GOPAL	76	90.5
13	HANGUP	80.5	94.75
14	HASH	63	92.75
15	HEMANT	73	95.25
16	HOLD	76.75	92.5
17	JAMES	71.5	96
18	JHONSON	75.25	93.25
19	JOHN	77.25	92.5

Sl. No.	Words	Recognition accuracy (in percentage)	
		Monophones based HMMs	Triphones based HMMs
20	KRISHNA	69.25	91
21	MAYANK	72.25	95.5
22	NIL	69	92
23	NINE	80.5	94
24	NITIN	69.5	95.75
25	OFF	74.5	93
26	ON	75.5	94.25
27	ONE	81.5	96.75
28	PHONE	83	97.75
29	PUNIT	63.75	90.5
30	RAJA	76.25	95.75
31	RAM	69.25	94
32	REDIAL	77.5	93
33	ROHIT	73.25	91
34	SEVEN	78.25	97.25
35	SIX	83.5	98
36	SPEAKER	75.25	92
37	STAR	71	93.25
38	STEVE	72	94.25
39	SUMIT	65.75	90.75
40	THREE	76.75	94.25
41	TWO	62.75	90.25
42	ZERO	78	93.25

Table 4: Word recognition performance using Monophones and Triphones HMMs.

Table 5 below shows the confusion matrix of the mostly mis-recognized words and also for the word "six" which gives good recognition result.

Words	Words misrecognized as (percentage of misrecognition)				
TWO	ONE (10.5)	THREE (4.5)	NINE (3.25)	ZERO (7)	FOUR (2.25)
GOPAL	ROHIT (3.25)	ANIL (4.5)	JAMES (1.75)	JOHN (2.25)	BIKASH (2.75)
PUNIT	NIL (3.25)	GOPAL (4.75)	HEMANT (2.75)	ROHIT (6.75)	AMIT (9.25)
SUMIT	AMIT (9.75)	PUNIT (4.5)	JOHN (2)	HEMANT (1)	NITIN (7.75)
KRISHNA	GOPAL (3.25)	NITIN (5)	JHONSON (4.25)	JOHN (4)	BIKASH (5.25)
ROHIT	JAMES (7.5)	PUNIT (2)	JAMES (1.75)	HEMANT (5.25)	AMIT (1.25)
EIGHT	NINE (8.25)	TWO (8.5)	FIVE (6.5)	THREE (2.75)	-
FLASH	HASH (15.25)	HOLD (8.5)	STAR (4)	-	-
HASH	FLASH (16.5)	HOLD (8.5)	HANGUP (4.75)	-	-
SIX	ONE (5.5)	SEVEN (9)	-	-	-

Table 5: Words that are confused with other words.

5. ANALYSIS AND DISCUSSIONS

From the above Table 4 it has been observed that for monophones based system, the word **hash** and **flash** gives low recognition score. This may be due to the fact that the pronunciation of both the words is very similar to each other. Also the words **two, Rohit, Punit** and **Sumit** gives poor results. From this we can infer that the words containing the dental sound /t/ can confuse the system. Sometimes the pruning of the words during labeling is not done properly and some part close to the boundary of the words gets removed. As a result of this, the models formed by that data is not properly trained and it gives low recognition score. So while cutting the word for labeling it is necessary to leave some silence region before and after each word in the data preparation stage.

The two Figures 2 and 3 shown below depicts the spectrogram of worst recognized word 'TWO' and the best recognized word 'SIX' found in the course of the experiments. It is clear from the plots that though the formants are well marked and steady for "two", there is less variations that can be captured and modeled by our system. On the other hand the spectrogram plot for "six" clearly shows numerous acoustic changes in the formant structure during the utterance. We can infer that this property of modeling the transitions is well captured by the triphones based HMMs.Though some earlier experiments show poor result for the word "six", in our experiment "six" gives very good result. This may be due to the good recording quality or may be the spotting and pruning of word is proper for "six".

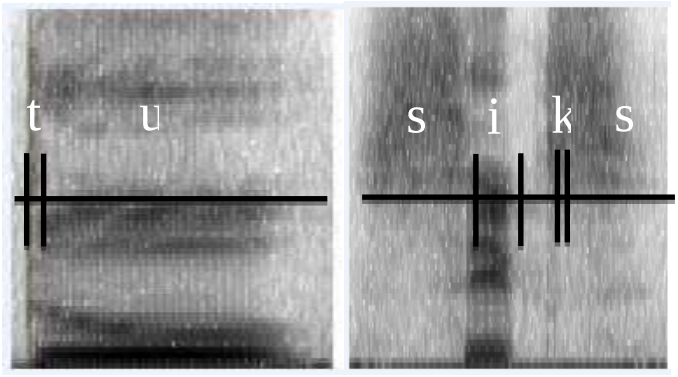


Figure 2. Figure 3.
Spectrogram of "Two". Spectrogram of "Six".

For triphones based recognition system it is observed that all the words give reasonably good result as compared to monophones based system. The monophones based recognition system gives accuracy of 74.11% while triphones based recognition system gives 93.77%. So, from the experiment we can clearly say that the triphones model gives much better results than the monophones-based model.

6. CONCLUSIONS AND FUTURE WORK

In this paper two telephone-based recognition systems are developed with the help of monophones and tri-phones using the Hidden Markov Toolkit (HTK). The system is able to

recognize the words spoken by speakers inside and outside the set for both continuous and isolated words. The whole experiment has been carried out in a normal room environment. The system gives good accuracy of 74.11% for monophones and 93.77% for triphones based models. The triphones based models perform far better than the monophones based models. Work is now underway to semi-automate the generation of the training models to deploy a speech recognition system at a short notice. The authors are also planning to update the computation of the feature vectors in *hcopy* to include new acoustic-phonetic features within the toolkit.

REFERENCES

[1] Shivendra Goel, J. B. Singh, Ashok Kumar Sinha, "Traffic Generation Model For Delhi Urban Area Using Artificial Neural Network", BIJIT - BVICAM's International Journal of Information Technology, Vol. 2, No. 2, 2010.

[2] R. K. Agrawal, Ashish Chaudhary, "Modified Incremental Linear Discriminant Analysis for Face Recognition", BIJIT - BVICAM's International Journal of Information Technology, Vol. 1, No. 1, 2009.

[3] Anil K. Jain, Robert P. W. Duin, Jianchang Mao, "Statistical Pattern Recognition: A Review",IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1,2000, pp. 4-37.

[4] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition",PTR Prentice Hall, 1993, ISBN 0130151572, 9780130151575.

[5] L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", Inequalities, Vol. 3, 1972, pp. 1-8.

[6] J**.** K. Baker, "The Dragon System - An Overview", IEEE Transactions on Acoustic Speech and Signal Processing, Vol. ASSP-23**,** No. 1, Feb. 1975, pp. 24-29.

[7] F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proceedings of IEEE,Vol. 64**,** April 1976, pp. 532-536.

[8] Vinciarelli A and Luettin J., "Off-line Cursive Script Recognition based on Continuous Density HMMs", Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, 2000, pp. 493-498.

[9] Md. Rafiul Hassan and Baikunth Nath, "Stock Market Forecasting using Hidden Markov Model: A New Approach", IEEE Proceedings of the 5th International Conference on Intelligent Systems Design and Applications (ISDA), 2005, pp. 192-196.

[10] Gupta, R., "Speech Recognition for Hindi", M. Tech. Project Report, Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India, 2006.

[11] Kuldeep Kumar, R. K. Aggarwal, "Hindi Speech Recognition System using HTK", International Journal of Computing and Business Research, ISSN (Online) : 2229-6166, Volume 2, Issue 2, May 2011. http://www.researchmanuscripts.com/PapersVol2N2/IJCBRVOL2N2P3.pdf

[12] Nguyen Hong Quang, Trinh Van Loan, Le The Dat, "Automatic Speech Recognition for Vietnamese using HTK System", IEEE International Conference on Computing and Communication Technologies Research, Innovation and Vision for the Future (RIVF), Hanoi, Nov. 2010, pp. 1-4.

[13] Bassam A. Q., Al. Qatab, Raja N. Ainon, "Arabic Speech Recognition using Hidden Markov Model Toolkit (HTK)", Information Technology International Symposium (ITSim), Kuala Lumpur, 2010, pp. 557-562. http://noble.gs.washington.edu/papers/htk-sequence.pdf

[14] P C Woodland, J J Odell, V Valtchev, S J Young, "Large Vocabulary Continuous Speech Recognition Using HTK", Proceedings of ICASSP 94 IEEE International Conference on Acoustics Speech and Signal Processing, Adelaide, 1994, pp. II/125-II/128

[15] HTK Official website on History of HTK page [Online] Available: http://www.htk.eng.cam.ac.uk

[16] A. N. Mishra, Mahesh Chandra, Astik Biswas, S. N. Sharan, "Robust Features for Connected Hindi Digits Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 4, No. 2, June 2011, pp. 79-90.

[17] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland. The HTK Book Version 3.4.1, 2009.

WORDS	PHONEMES	IPA
AMIT	ae m ih t	/əmit/
ANIL	ae n ih l	/ənil/
ARUN	ea r ax N	/ərun/
BIKASH	b ih k ax SH	/bika:ʃ/
CALL	k ao l	/kɔl/
DEBANJAN	d b ae n jh ae n	/deba:ndʒən/
DIAL	d ay ax l	/daɪl/
EIGHT	ey t	/eɪt/
FIVE	f ay v	/faɪv/
FLASH	f l ae sh	/flæʃ/
FOUR	f ao r	/four/
GOPAL	g ow p l	/gopa:l/
HANGUP	hh ae ng ah p	/ˈhæŋˌʌp/
HASH	hh ae sh	/hæʃ/
HEMANT	hh eh m ae n t	/hemənt/
HOLD	hh ow l d	/hould/
JAMES	jh ae m eh s	/dʒeɪmz/
JHONSON	jh oh n s ah n	/ˈdʒɒnsən/
JOHN	jh oh n	/dʒɒn/
KRISHNA	k r iy s n ax	/krɪʂna/
MAYANK	m ey ey ae ng k	/məjɔ̃k/
NIL	n ih l	/nɪl/
NINE	n ay n	/naɪn/
NITIN	n ih t ih n	/nitin/
OFF	oh f	/ɒf/
ON	oh n	/ɒn/
ONE	w ah n	/wʌn/
PHONE	f ow n	/foʊn/
PUNIT	p uh n ih t	/pʊnit/
RAJA	r aa jh ax	/rɑ:dʒa/
RAM	r ae m	/ra:m/
REDIAL	r iy d ia l	/riˈdaɪəl/
ROHIT	r ow hh ih tH	/rɔɦit/
SEVEN	s eh v n	/ˈsɛvən/
SIX	s ih k s	/sɪks/
SPEAKER	s p iy k ax r	/ˈspikər/
STAR	s t aa r	/stɑr/
STEVE	s t iy v	/stiv/
SUMIT	s ah m ih t	/sumit/
THREE	th r iy	/θri/
TWO	t uw	/tu/
ZERO	z ia r ow	/ˈzɪərou/

Table 6: The phonetic breakup and IPA representation of all the words.

Performance Evaluation of Superscalar Processor Architecture Through UML

Taskeen Zaidi[1] and **Vipin Saxena**[2]

Submitted in May 2012, Accepted in October 2012

Abstract - In the current scenario, most of the applications are based upon graphical user interface and dependent upon the object-oriented technology. Software Industries are interested to convert old structured based softwares into object-oriented based softwares and also to reduce the lines of the code of application for reduction in the execution time of application. Therefore, it is a big challenge to reduce the execution time of the application based upon the object-oriented technology. The present work deals with the reduction of execution time for the superscalar machine by the use of object-oriented approach. A well known modeling language i.e. Unified Modeling Language (UML) is used to model the superscalar pipeline architecture. UML class and sequence models are designed before computations of the execution time and computed results are depicted in the form of tables and graphs. The comparisons are also made by taking the two object-oriented programming languages.

Index Terms - Superscalar pipeline architecture, performance evaluation, class model, sequence model and unified modeling language.

1. INTRODUCTION

Pipelining is one of the important techniques which have been implemented to improve the performance of a processor. It allows the concurrent execution of several instructions. A task or program or process is divided into sequence of subtasks and each task is executed by a specialized hardware stage which operates concurrently with other stage in pipeline. There are several categories of pipeline like arithmetic pipeline, instruction pipeline, memory access pipeline and superscalar pipeline. Superscalar pipeline architecture can start two or more instructions in parallel in one core, and independent instructions may get executed out-of-order. For parallelism, scalabilityand programmability, [1] is an important release which describes these aspects with increasing system resources and accordingly to parallel, vector and scalar instructions. Mano [2] describes the computer organization and design as well as programming using basic components.

Patterson and Hennessey [3] covers the most fundamental areas of computer architecture including recent technologies, like multicores and multiprocessors.

The depth treatment with the implemented details of pipelined processors and memory systems; the "micro architecture" of the modern computers and microprocessors by exploring the techniques for solving design problems inherent in computers with high level concurrency as the demand for a memory system with low latency and high bandwidth are described by Cragon and Saini [4,5].

Unified Modeling Language (UML) is a general purpose modeling language which is used to model various kinds of the research problem widely accepted by the software professionals and created by Object Management Group (OMG [6,7] and development stages are well explained by Booch et al. [8]. The fundamentals of UML using hands-on projects, drills and mastery checks which illustrates how to read, draw, and use this visual modeling language to create clear and effective blueprints for software development projects are explained by Roff [9]. UML is also used to model the concurrent distributed and real time applications which help the researchers to leverage the powerful flexibility and reliability of the system. UML also helps the designers at every stage of the analysis and design process and offers exceptional insight into dynamic modeling, concurrency and distributed applications designing and performance analysis of real time designs [10]. By using distributed computing, the performance of processors for different object-oriented software system framework has been measured by Saxena et al. [11]. They have chosen two types of object-oriented software system frameworks C# based on Microsoft.NET framework and Visual C++ based on Microsoft Foundation Classes (MFC) and computed the performance of these two object-oriented languages. The UML modeling for instruction pipeline design by two techniques i.e. data forwarding and without data forwarding are explained by Saxena and Raj [12]. The modeling and specification of floating point numbers are implemented by Boldo et al. [13]. It extends an existing tool for the verification of C programs, with the new notations specific to the floating point arithmetic. It also provides a way to perform the full formal proof by use of COQ proof assistant and an open framework which is implemented to other floating point models. But the main limitation is that it is applicable only to programs using basic. The IEEE standard is the most widely used standard for floating point and arithmetic representation. It is implemented on most of Central Processing Units (CPU's) and Floating Points Units (FPU's); explained with basic and extended floating point number formats, operations such as add, multiply, divide, square root, etc. It is also used to implement the conversion between integer and floating point formats, but, it does not specify the decimal strings and integers, interpretation of NAN's and conversion of binary to decimal to and from extended format [14]. Saxena and Shrivastava [15] have attempted to increase the performance of arithmetic

[1, 2] *Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow (U.P.), INDIA*
E-mail: [1]taskeenzaidi867@gmail.com and [2]vsax1@rediffmail.com

pipeline especially for floating point computations after designing the complete UML model of static arithmetic pipeline design. They presented UML diagrams to model the system architecture and timing behavior. Saxena and Shrivastava [16] also presented floating point computations by using nonlinear arithmetic pipelining for instruction coupled on Visual C++ and Visual C#. The computations are performed inside a loop by varying the number of repetition of terms for getting their sum. In the current scenario, distributed computing approach is most popular approach and in this regards, a comparative study of the distributed computing paradigms is presented in [17]. Quality of services is one of the major issue for the distributed computing applications and these are described by Mohan et al. [18] for the process centric development. The design patterns for the service oriented architecture implementation are described by Tere and Jhadav[19].

In the present work, UML is used to model class and sequence diagrams for superscalar pipeline architecture which can execute two or more instructions in parallel and authors evaluated the performance of the two object-oriented languages like Visual C++ and Visual C# and some of the important observations are recorded in the form of table and graphs.

2. BACKGROUND

2.1 Process Definition

Let us first explain the process which is considered as a program which is to be executed. It can be defined as a unit of work in modern time sharing systems. For defining the process processing element is needed to be defined as stereotype and is used to handle some modeling elements based on UML base classes. A UML Class for process is shown in figure1 and is identified by its own identification number represented as Process-id. The other attributes are Process-size for the size of a process; Process_in_time and Process_out_ time are for start at out time of the process. The attribute Process_priority controls the priority of the incoming process. These attributes work on the operations like Process_create(), Process_delete, Process_ update, Process_ join, Process_suspend, and Process_synchronize. The visibility modes along attribute and operation are also shown in the figure. A stereotype processing unit is also depicted in figure 2, the instance and multiple instances of class Process class are shown in Figures 3(a) and 3(b), respectively. The process may consist of segments of code whose identification numbers are generated; recorded into a list and granted processing unit as per the priority of that segment code behaving as a process. The segments may be synchronized with the processing unit as per the time of completion of that segment; therefore, multiple instances of a process are shown in figure 3(b).

2.2 Thread

A thread is defined to control a block of code that runs concurrently with other threads within same process. It is a sequential flow of instructions and it is considered as lightweight process. It is easily handled in object-oriented way. Threads run simultaneously in process and can access the same object to implement their functionality.

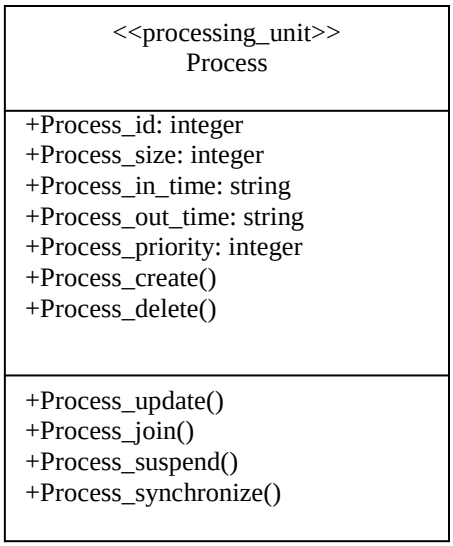


Figure 1: UML class diagram of a process

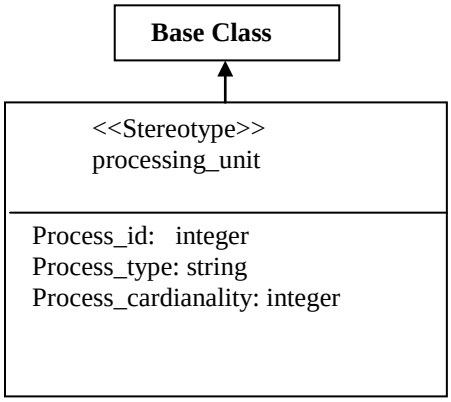
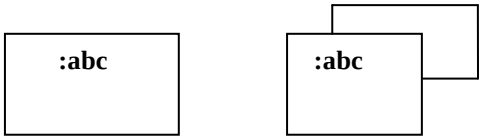


Figure 2: UML class for processing unit



3(a) **3(b)**

Figure 3a: Single instance, Figure 3b: Multiple instances

In the current scenario, most of the window based applications are based upon the thread concept as system supports synchronization of sub tasks of a process. Threads are initialized and after the use these are automatically destroyed, therefore, it has a life cycle. Object-oriented representation of thread is shown below in figure 4, in which it is identified by an attribute called as Thread_id. The other attributes associated with thread and thread operations are also shown below in the figure 4.

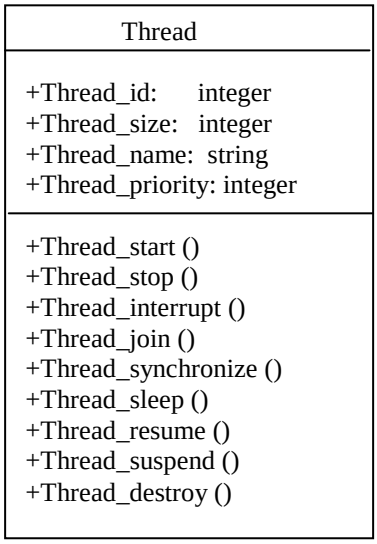


Figure 4: UML Class Diagram of a Thread

2.3 Superscalar Processor Architecture

Superscalar processor architecture has a versatile design with the two pipelines and it can issue the instructions per cycle, if there is no resource conflict and no data dependence problem. Both pipelines have four processing stages namely fetch, decode, execute and store.

Each pipeline has its own fetch, decode, execute and store unit. The two store units can be dynamically used by the two pipelines, depending upon its availability at particular cycle. It has four functional unit adder, multiplier, logic and load unit. These all functional units are shared by pipelines on dynamic basis. There is a lookahead window with its own fetch and decoding logic. Lookahead window is used in case of out of order instruction to achieve better pipeline throughput.

3. UML MODELING FOR SUPER SCALAR PIPELINE DESIGN

3.1 UML Class Diagram

The figure5 shows the architectural model of superscalar processor. The class process interacts directly with PEC which executes the assigned task. The PEC controlled the process by exchanging message between classes processor and memory. The processor class has two cores i.e. Core1 and Core2 and each core has many components which help in process execution as shown in figure.

In this figure, class L2_cache is shared by two cores and caches instruction through the class I Cache whereas D_cache caches the data, which itself is subclass of L1_cache. The class ALU computes integer arithmetic and logical operations; FPU is used for floating point operations as shown in figure. SU is used for storing the outputs. FPU class contains four classes as namely Adder, Multiplier, Logic and Load unit.

3.2 UML Sequence Diagram

The UML Sequence diagram represents the dynamic behavior of system in which objects are interacted with the help of message communications. The vertical line shows the life line of object or dynamic representation of system, a UML sequence diagram is shown in figure 6 for process execution in Superscalar pipeline architecture.

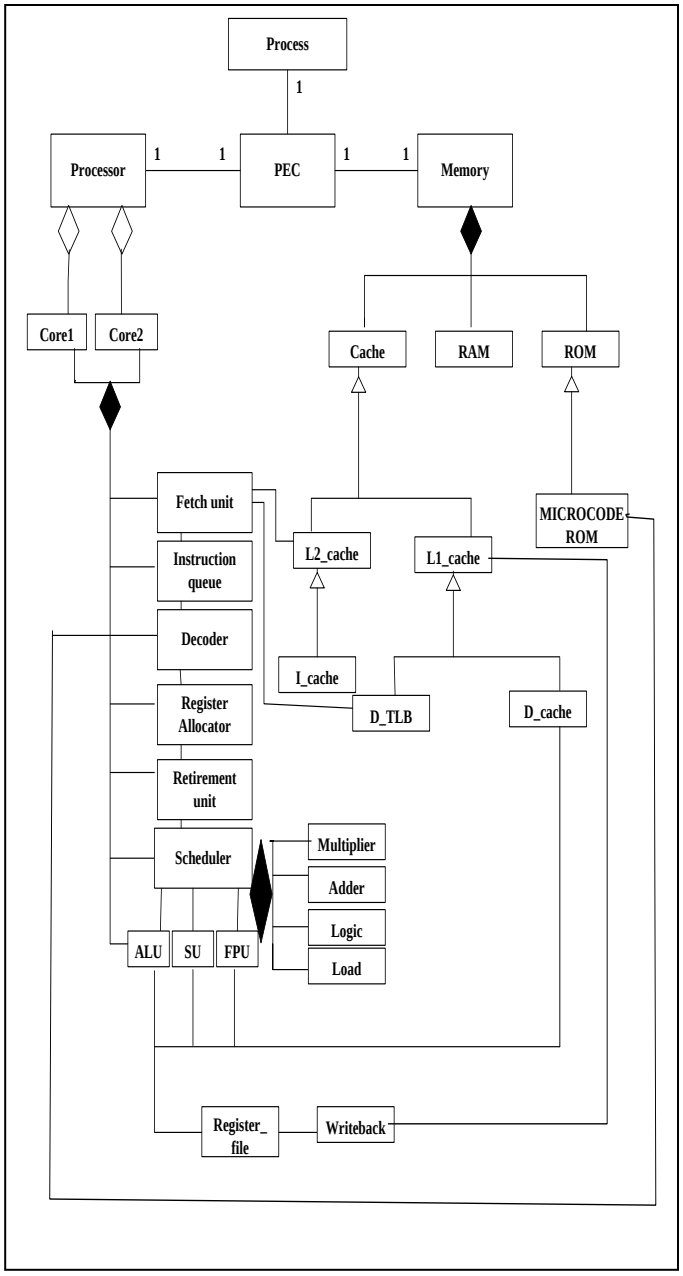


Figure 5: UML class diagram for superscalar process

The processor executes the instructions fastly through execution pipelining, which execute multiple instructions at same time. The instruction fetched, decoded and finally goes to PEC where instructions executed and results store in Registerfile and then Writeback.

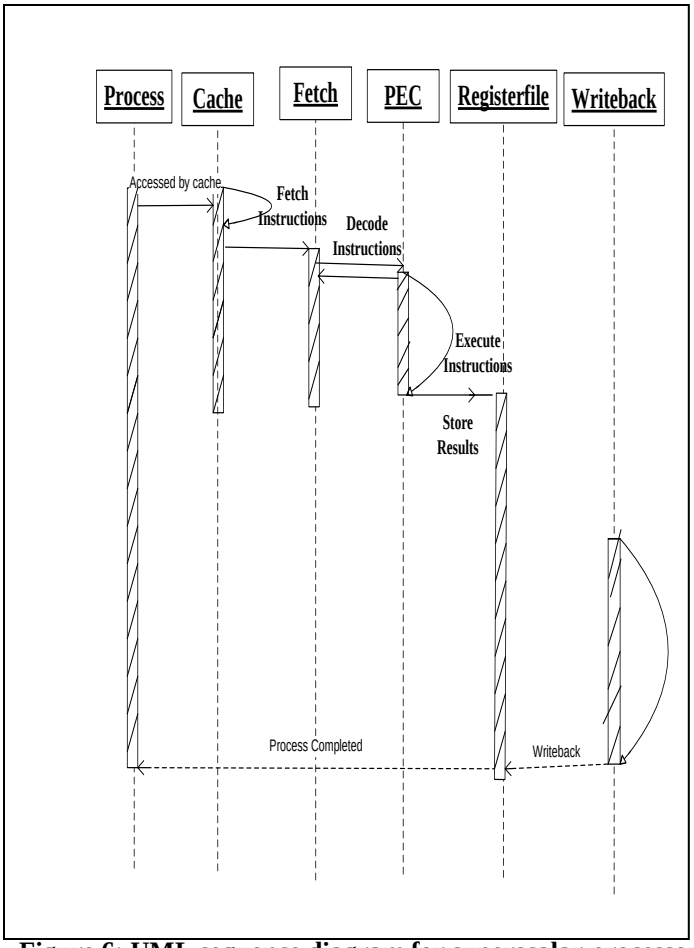


Figure 6: UML sequence diagram for superscalar processor

4. EXPERIMENTAL STUDY

On the basis of above object-oriented design, let us consider the two object oriented languages i.e. VC++ and VC# which work on the .Net platform. For these two programming languages, a relative performance of the superscalar processor is computed. In the computation, let us consider N are the independent instructions which can be executed in parallel through pipeline method and k is taken as the time required to execute instructions through m pipeline simultaneously, then the ideal time required by scalar base machine is

$T(1, 1) = k + N - 1$ ……. (i)

The ideal execution time is computed by

$T(m, 1) = k + (N-m)/m$…….. (ii)

The computations for ideal execution time are done by taking lines of code varying from 10^2 to 10^5 and these instructions are considered by increasing the size of loop. Execution time is computed by taking average of five runs and results are depicted in the table1. As expected lines of code are increasing, execution time is also increasing but if one compares VC++ andVC#, for long computations in milliseconds, it is observed that VC++ takes lesser time in computation than VC#. These results are also graphically represented infigures 7 and 8 given on next page for 10^2, 10^3 and 10^4,10^5 lines of code (LOC), respectively.

5. CONCLUSION

From the above work, it is concluded that UML is powerful modeling language accepted by software Professionals and also used to represent hardware architecture problems. For the long computations, software professionals are facing the problems for selection of best object oriented Programming language which works well on any kinds of processor architecture.

Therefore, superscalar processor architecture is modeled by the use of UML classes and experimental results are performed by taking two object-oriented programming language like Visual C++ and Visual C# and concluded that Visual C++ is better in comparison to Visual C# as one is performing the long computations.

REFERENCES

[1]. Hwang, K., Advanced Computer Architecture: Parallelism, Scalability, Programmability, Fourteenth Reprint, Tata McGraw-Hill Edition, ISBN-0-07-053070-X-2007.

[2]. Mano Morris, M., Computer System Architecture, Third Edition, Prentice Hall of India Pvt Ltd. ISBN-978-81-203-0855-8, 2007.

[3]. Patterson, A. David and Hennessy, L. John, Computer Organization and Design: The Hardware/Software Interface, Morgan Kaufmann Publishers Elsevier Inc., 2005.

[4]. Cragon, G. Harvey, Memory Systems and Pipelined Processors, Narosa Publishing House, New Delhi, 1998.

[5]. Saini, A., "Design of the Intel Pentium TM Processor", Intel Corporation, IEEE, and Available: http://ieee xplore.ieee.org/stamp/stamp.jsp?arnumber=00393370 (Accessed on 14th March 2012).

[6]. OMG (2001),"Unified Modeling Language Specification", Available online via http://www.omg.org.

[7]. OMG (2002), "XML Metadata Interchange (XML) Specification", Available online via http://www. omg.org.

[8]. Booch, G., Rambaugh, J., and Jacobson, I., The Unified Modeling Language User Guide, Twelfth Indian Reprint Pearson Education, 2004.

[9]. Roff, T., UML: A Beginner's Guide, Tata McGraw–Hill Edition. Fifth Reprint, 2006.

[10]. Gomaa, H., "Designing Concurrent, Distributed and Real Time Applications with UML", Proceedings of the 23rd International Conference on Software Engineering (ICSE'01), IEEE Computer Society, 2001.

[11]. Saxena, V., Arora, D., and Ahmad, S.; "Object Oriented Distributed Architecture System through UML", IEEE International conference on Advanced in Computer Vision and Information Technology (ACVIT-07), Nov. 28-30, ISBN 978-81-89866-74-7, pp.305-310,2007.

[12]. Saxena, V. and Raj, D., "UML Modeling for Instruction pipeline Design", World Conference on Science, Engineering and Technology (WCSET,2008), www .waset. org/ pwaset (Acessed on 16 NOV,2011).

[13]. Boldo, S. and Filliatre, J.C., "Formal Verification of Floating Point Programs", 8th IEEE Symposium on Computer Ari-thmetic (ARITH '07), pp.187-194 . Availabale: http :/ / www. computer.org (Accessed on 16 Nov, 2011).

[14]. Lopez, G., Taufer, M., and Teller, P.J., "Evaluation of IEEE 754 Floating-Point Arithmetic Compliance across a wide range of Heterogeneous Computers", Proceedings of the 2007 Richard Tapia Celebration of Diversity in Computing Conference, October 2007 , Orlando, Flor-ida ,USA. Available:http://gcl.cis.udel.edu/publication/ conferences/ 007tapia_mlopez.pdf (Accessed on 16 Nov, 2011).

[15]. Saxena, V. and Shrivastava, M., "UML Modeling of Static Arithmetic Pipeline Design", The ICFAI University Press Vol. 7(1), pp.22-31, February 2009.

[16]. Saxena, V. and Shrivastava, M., "Performance Evaluation of Non-Linear Pipeline through UML", International Journal of Computer and Electrical Engineering,Vol.2, No.5, pp.860-866, October, 2010.

[17]. Kumar, H. and Verma, A.K., "Comparative study of Distributed Computing Paradigms", BIJIT – BVICAM's International Journal of Information Technology, Vol. 1(2), Dec. 2009.

[18]. Mohan, K.K., Srividya,A., Verma, A.K. and Gedela, R.K., "Process Centric development to Improve Qos in Building Distributed Applications", BIJIT – BVICAM's International Journal of Information Technology, Vol. 1(1), July, 2009.

[19]. Tere, G.M. and Jhadav, B.T., "Design Patterns for successful Service Oriented Architecture Implementation", BIJIT – BVICAM's International Journal of Information Technology, Vol. 2(2), Dec. 2010.

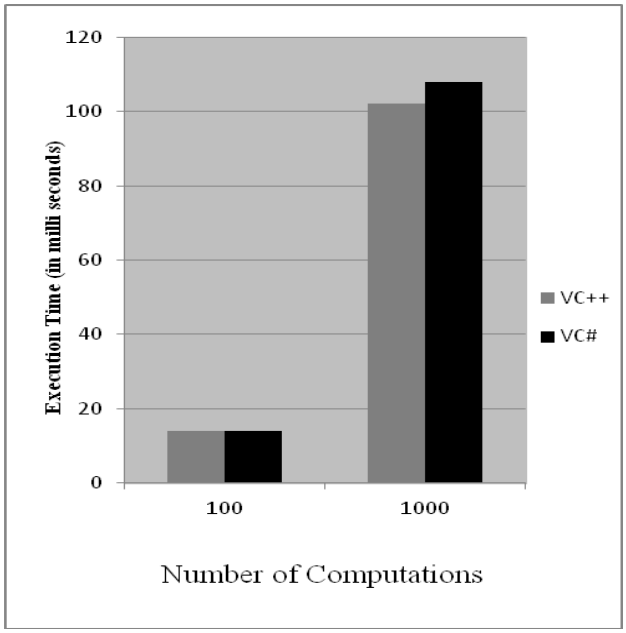


Figure 7: Comparisons for 10^2 and 10^3 Lines of Code

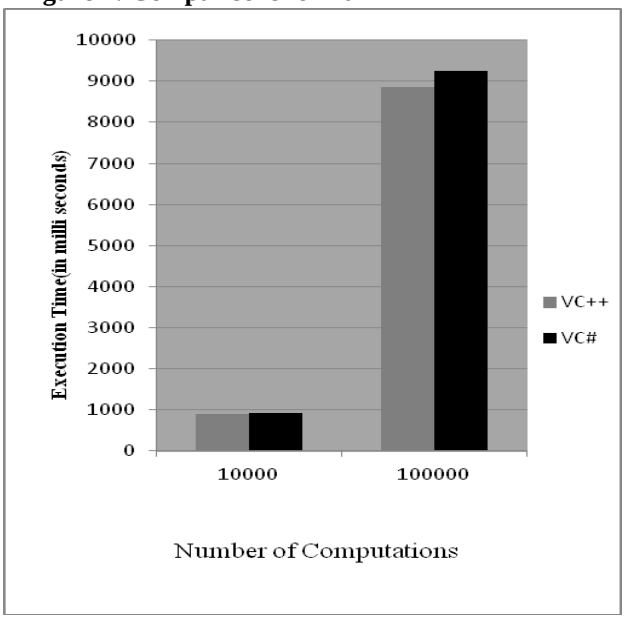


Figure 8: Comparisons for 10^3 and 10^5 Lines of Code

	VC++				VC#			
Lines of Code	10^2	10^3	10^4	10^5	10^2	10^3	10^4	10^5
Execution Time (in Milli Seconds)	14.05	92.005	889.0005	8952.00005	14.05	108.005	920.0005	9233.00005
	14.05	109.005	889.0005	8967.00005	14.05	108.005	936.0005	9170.00005
	14.05	108.005	874.0005	8780.00005	14.05	108.005	920.0005	9264.00005
	14.05	93.005	890.0005	8796.00005	14.05	108.005	920.0005	9249.00005
	14.05	108.005	905.0005	8812.00005	14.05	108.005	936.0005	9280.00005
Average Execution Time	14.05	102.005	889.4005	8861.40005	14.05	108.005	926.40005	9239.20005

Table 1: Ideal execution time for superscalar processor

Survey of Energy Computing in the Smart Grid Domain

Rajesh Kumar[1] and **Arun Agarwala**[2]

Submitted in September 2012, Accepted in March 2013

Abstract - Resource optimization, with advance computing tools, improves the efficient use of energy resources. The renewable energy resources are instantaneous and needs to be conserve at the same time. To optimize real time process, the complex design, includes plan of resources and control for effective utilization. The advances in information communication technology tools enables data formatting and analysis results in optimization of use the renewable resources for sustainable energy solution on smart grid.

The paper presents energy computing models for optimally allocating different types of renewable in the distribution system so as to minimize energy loss. The proposed energy computing model optimizes the integration of renewable energy resources with technical and financial feasibility. An econometric model identifies the potential of renewable energy sources, mapping them for computational analysis, which enables the study to forecast the demand and supply scenario. The enriched database on renewable sources and Government policies customize delivery model for potential to transcend the costs vs. benefits barrier. The simulation and modeling techniques have overtaken the drawbacks of traditional information and communication technology (ICT) in tackling the new challenges in maximizing the benefits with smart hybrid grid. Data management has to start at the initial reception of the energy source data, reviewing it for events that should trigger alarms into outage management systems and other real-time systems such as portfolio management of a virtual hybrid power plant operator. The paper highlighted two renewable source, solar and wind, for the study in this paper, which can extend to other renewable sources.

Index Terms - Energy Computation, Energy Mapping, Techno-Economical feasibility of Renewable Energy, Renewable energy model, Energy Efficiency

1. INTRODUCTION

Supervisory Control and Data Acquisition (SCADA) systems for control on hybrid sources of energy have two components: Energy Management Systems (EMS) and Distribution Management Systems (DMS). A hybrid EMS/DMS system requires higher level security analysis functions such as state estimation and contingency analysis for EMS and feeder voltage and loss optimization for DMS systems.

[1, 2] *IDDC, Indian Institute of Technology (IIT) Delhi, New Delhi*
E-mail: [1]rajeshkr38@nic.in and [2]agarwala@iddc.iitd.ernet.in

Energy Distributive system model for adequate accurate predictive analysis plays important role for sustainable country energy resources, in consideration of all influential factors in energy generation and distribution. For prediction purposes the important parameters are geographical location, seasonal influence, effect of climate change and state or area concession. Renewable energy potential for mitigation action on climate change reported by IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation run on four models namely, IEA-WEO2009-Baseline, ReMIND-RECIPE, MiniCAM-EMF22, ER-2010 for potential scenarios [1]. With the abundant data and relative economic indicator, energy prediction is performed with close loop predictive system based on a timing algorithm. The energy economics in free trade market like India, where the peak load varies abruptly due to season and community demand, its hourly prediction model is more useful.

The energy prediction model proposed in this paper has a large scope to take on innovative role for country's growth in the energy security regime. Scientists are working on commercial application of energy modeling. The computation and mapping tool is unique in that city planners and government to integrate renewable energy on the grid. The tool is helpful for planning new substations and infrastructure in the ever-growing city.

The study optimizes the integration of the various renewable energy resources with financial feasibility. The model overcomes the constraints like hourly available sources, the voltage limits, the feeders' capacity, and the discrete size of the available distributive generation and distribution units. This paper addresses the following issues:

- Power grid planners need to account for the impacts brought by different kinds of energy sources like power factor, hybrid energy voltage, load management programs, energy efficiency, high renewable energy penetrations, and energy storage.
- Evaluation of the cost/benefit of the different technologies.
- Setup a planning tool to run a base case and a comparable case that has a new technology implemented.
- Generate a cost effective/optimal expansion plan.

The paper is organized to present the modeling approach in computing complex energy scenario of demand and supply. The paper has computation and algorithms for modeling results of solar and wind renewable resources for conclusion and recommendation.

2. REVIEW OF ENERGY MODELS

Energy is a vital input for social and economic development of the community and the state. In technology driven economies the demand for energy in agricultural, industrial and domestic activities has increased remarkably, especially in emergent

countries, which also increases greenhouse gases. The cost economics of energy forces the use of renewable energy sources more effectively, i.e. energy which comes from natural resources and is also naturally replenished. The dependability of renewable energy resources on the climate enhance the need for complex design, planning and control optimization methods [7].

Power system planning involves planning of generation, transmission, and distribution systems. Generation planning begins with mid-term (months to several years) and long-term (several years to 10 years) load forecasts because generation expansion often requires 2 to 10 years to complete. When load forecast is available, reliability evaluations will be the next step to assess where and when to install the new generation. Finally, economic evaluations are performed to determine the optimal generation expansion planning. Accurate load forecast leads to an economical capacity expansion plan that meets reliability requirements [10].

Leading vendors of power system planning tools are: Multi Area Production Simulation Software program (MAPS) from General Electric (GE), Plexos for Power Systems from Energy Exemplar, GridView from ABB, and PROMOD from Ventyx [20].

National Instruments Labview and the Labview Control Design and Simulation Module can be used to simulate a full wind turbine system, including the turbine, mechanical drive train, generator, power grid and controller. AROMA model method has been employed in a predictive model [11].

HOMER is a computer model developed by the U.S. National Renewable Energy Laboratory (NREL) to assist in the design of micro-power systems. HOMER finds the feasibility of the system by assessing whether it can adequately serve the electric and thermal loads through an hourly time series simulation over one year. It also estimates the life-cycle cost of the system, which is the total net present cost of installing and operating the system over its lifetime.[6]

The RET Screen Plus Performance Analysis Module can be used worldwide to monitor, analyse, and report key energy performance data to facility operators, managers and senior decision-makers.

The MARKAL model uses an integrated energy system optimization framework that enables policymakers and researchers to examine the best technological options for each stage of energy processing, conversion, and use. This modeling framework was used to represent a detailed technological database for the Indian energy sector with regard to energy resources (indigenous extraction, imports, and conversion) as well as energy use across the five major end-use sectors (agricultural, commercial, residential, transport, and industrial)[6].

LINGO is a comprehensive tool designed to make building designs. It can solve Linear, Nonlinear (convex & nonconvex/Global), Quadratic, Quadratically Constrained, Second Order Cone, Stochastic, and Integer optimization models efficiently. [26]

Renewable energy potential for mitigation action on climate change reported by IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation run on four models namely, IEA-WEO2009-Baseline, ReMIND-RECIPE, MiniCAM-EMF22, ER-2010 for potential scenarios. There is enormous variation in the detail and structure of the models used to construct the scenarios. Many authors have, in the past, attempted to categorize models as either bottom-up or top-down [1].

These models have constraints including hourly available sources, the voltage limits, the feeders' capacity, the maximum penetration limit, and the discrete size of the available DG units, with in the legal constraints applicable [14].

3. ENERGY COMPUTATION AND MAPPING MODEL

Accurate predictive analysis influential factors in energy application have been taken into consideration during design of energy model in this paper[20]. Taking the collected abundant data and related economic indicator model is accepted by international trade standards as the base, further strict calculation can lead to relative economic indicators. For prediction, full consideration of seasonal influence on renewable energy application must be considered with a closed-loop predictive system formed based on timing algorithm, to make the predictive model able to provide perfect prediction in the light of varied data [4][5].

The computing and mapping is addressed to the energy demand and the potential energy resources. Available data are collected based on a particular sampling procedure on field works and survey in 2007 and continued in 2010. This mapping is expected capable of informing accurate data about renewable energy diversification distribution all over the province. The global data sets and analytical tools at National Renewable Energy Laboratory (NREL) and for India specific at Centre for wind Energy Technology (C-WET) and Indian Metrology Department (IMD) permit modeling of wind and solar radiation resource predictions [19].

The proposed model for energy computing and mapping is extensions of AROMA model in Indian conditions. ARIMA prediction algorithms model by Peng Chen et-al provide a reliable base for popularization of renewable energy source application in building construction, key technologies, which include the multi-level system framework, functional modules, database design [11]. The present study is focused on two types of renewable i.e. Solar and Wind [14].

The proposed method has been employed in predictive model have higher accuracy of time sequence. In this study, the auto regressive integrating moving average model, will be study and analyse for the adoption within considered condition [9]. Then predictive monitor will done by employing model, plus comparison of predictive monitoring results and historical data, so as to achieve even better predictive monitoring results. The formula used in ARIMA model is described as the following,

$X_t = \psi_1 X_{t-1} + \ldots\ldots + \psi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \ldots. + \theta_q a_{t-q}$ ---- (1) which $\psi_1 \ldots \psi_p$ is the autoregressive coefficient, p is autoregressive order, $\theta_1 \ldots \theta q$ is moving average coefficient, q is moving average order, { a_t …}is noise sequence, X_t is the original data sequence, y_t is a stationary sequence formed through d times differential [11].

3.1 Solar Energy Mapping

Solar radiation assessment stations provide measurements of global solar radiation available and this methodology is called directly and for locations where the data was not available, indirect methods were used [19]. The indirect methods are as follows;

- From extra-terrestrial radiation, allowing for its depletion by absorption and scattering by atmospheric gases, dusts, aerosols and clouds. This is theoretically based and requires some approximation of the absorbing and the scattering property of the atmosphere.
- From other meteorological elements, such as duration of sunshine and cloudiness using regression technique. This method is empirical based, and the form usually used involves actual and potential hours of sunshine, which gives the regression constants for global and diffused solar radiation at the particular location or site.

The solar energy data is collected, documented and analysed by Ministry of New & Renewable Energy (MNRE) and Indian Metrological Department, MNRE has published the solar radiation potential map for India [8]. The solar energy is converted into useful energy with two techniques explained here.

A) Photovoltaic Power

Solar energy photovoltaic power is the direct solar energy utilization form with non-pollution, effective and easy power generation which can be either independent running or parallel running. The independent running of solar energy photovoltaic power generation system requires battery as the energy storage device, chiefly adopted in remote areas without power grid and dispersedly populated areas. But, the whole system is rather costly. In areas where power grid is available, the parallel running shall not only lower down the cost greatly, but also highly efficient with a friendly environment features. Systematically collect the global solar radiation, the solar radiation capacity and the parameter of effective radiation surface area of the solar cell array for evaluation of solar energy photovoltaic efficiency applied in buildings, the economic indicators shall be calculated as follows:

1. The global solar radiation I_R obtained from the surface of the solar cell array
2. The energy in the form electrical voltage and current produced by solar cell array is P_V
3. The inverter loss during conversion to usable energy L
4. Substituted quantity S_{PV} of conventional energy power conversion

As the important data of evaluating solar energy photovoltaic power efficiency applied in buildings, based on the above-listed economic indicators, to obtain the solar energy photovoltaic power model economic indicators assemble solar energy photovoltaic power

E_{SPV} { I_R, P_V, L, S_{PV} } --- (2)

The solar photovoltaic (PV) market saw another year of extraordinary growth. Almost 30 GW of new solar PV capacity came into operation worldwide in 2011, increasing the global total by 74% to almost 70 GW as shown in figure 1 [16].

B) Solar Thermal

The solar thermal system is another form of solar energy utilization. The system is to collect solar radiation energy through a device named heat arrester to heat exchanger. Such installation is presently the most economical and technically mature product which is already commercialized [3][21]. While evaluating efficiency of the solar energy arrester, the following five economic indicators shall be considered:

1. Solar energy assurance factor φ
2. Solar energy heat collecting system efficiency η_1
3. Heat exchanger efficiency η_2
4. Useful heat quantity of solar heat collecting system Q_{uf}
5. Substitution quantity of conventional energy sources S_{PV}

Based of the above-listed five economic indicators, the assemble indicator of solar water heating is thus obtained as

E_{ewh} { φ, η_1, η_2, Q_{uf} , S_{PV} } --- (3)

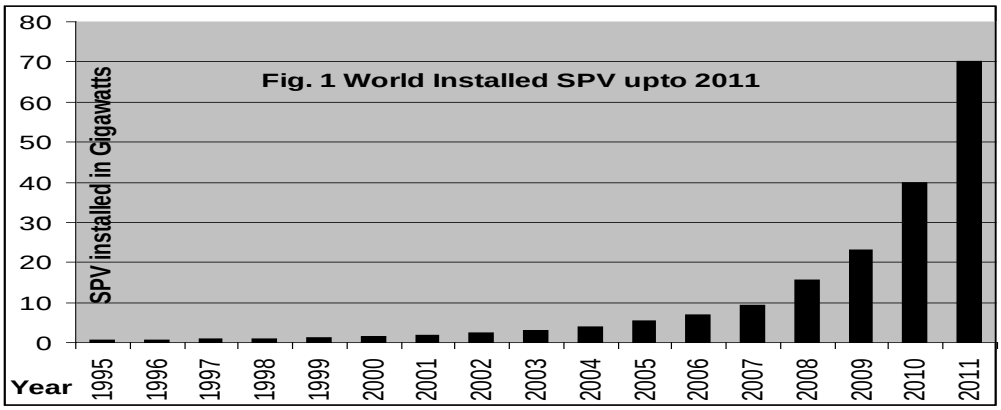


Figure 1: World installed SPV upto 2011

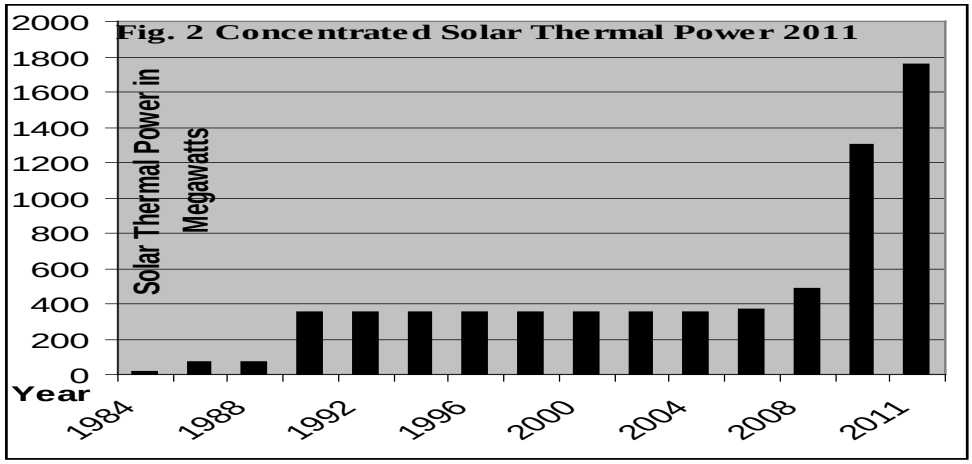


Figure 2: Concentrate Solar Thermal Power 2011

The concentrating solar thermal power (CSP) market continued its steady growth in 2011. More than 450 MW of CSP was installed, increasing total global capacity by 35% to nearly 1,760 MW [16]. The market was down relative to 2010, but significant capacity was under construction n at year's end. Over the five-year period of 2006–2011, total global capacity grew at an average annual rate of almost 37%. (See Figure 2.)

3.2 Wind Energy Computation

Wind energy potential is calculated based on the wind data on annual average wind speed. Annual average wind velocity data for wind-monitoring stations across Indian states are collected by the India Meteorological Department (IMD). To analyze variations across seasons, data was grouped season wise as summer (February–May), monsoon (June–September)and winter (October–January). Season wise wind velocity and standard deviation are computed for wind-monitoring stations. GIS is used for mapping wind resources spatially and to quantify and analyse temporal changes. Based on these, GIS thematic layers are generated, which would help in assessing the variability. The map helps to identify the most and the least suitable potential areas for harnessing wind energy.

The wind turbines power curve is defined as the power output of the machine as a function of wind speed. The behavior of the output power of the machine is generally dependent on four characteristic parameters. It is assumed that power generation starts at the cut-in wind speed V_C (m/s), that the output power increases as the wind speed increases from to the rated wind speed V_R (m/s), and that a constant value of the output power, namely the rated power P_R (kW), is produced when the wind speed varies from V_R to the cut-out wind speed V_F (m/s), which is the maximum wind speed value at which the turbine can correctly work.

The linear wind model assumes a linear (affine) dependence (within the interval [V_C & V_R]) of the wind turbine power output, P^t , on the current wind speed at the hub height V^t

As t=0 ,….. T-1, being T the time horizon in hours. In detail:

$$P^t = \begin{cases} 0 & V^t < V_C \\ P_R(a+bV^t) & V_C \le V^t \le V_R \\ P_R & V_R \le V^t \le V_F \\ 0 & V^t > V_F \end{cases} \quad T=0, \ldots, T-1 \qquad (2)$$

It should be observed that wind speed V^t in (2) is that corresponding to the wind turbine hub height, H_{hub}, Since, in general, wind speed data can be measured or forecasted with reference to a height H_{data}that is different from the hub height, it is necessary to use an equation relating the wind speed at hub

height with the wind speed V^{t}_{data} at H_{data} , taking into account the surface roughness length, which is a parameter that can be estimated on the basis of the land use at the wind farm location.

$$Ew\{ P_{R,} V_{C,} V^{t} , V_{F,} H_{hub} \} \quad (3)$$

During 2011, an estimated 40 GW of wind power capacity was put into operation, more than any other renewable technology, increasing global wind capacity by y 20% to approximately 238 GW as shown in figure 3.

Around 50 countries added capacity during 2011 and enhanced power capacity more than 10 MW in 68 countries and out of these 22 have cross 1 GW capacity. The top 10 countries account for nearly 87% of total capacity. Over the period from end-2006 to end-2011, annual growth rates of cumulative wind power capacity averaged 26% [16].

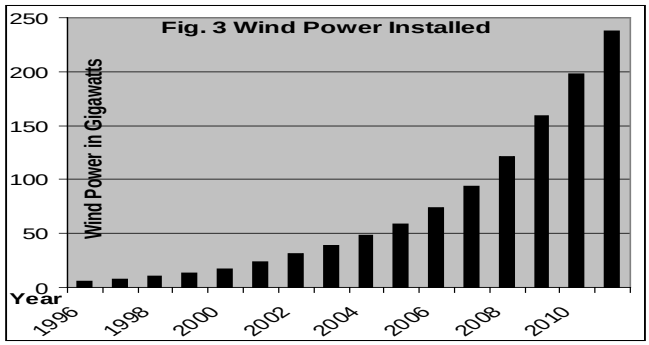


Figure 3: Wind Power installed

4. ECONOMICS OF ENERGY COMPUTATION

To achieve the accurate predictive analysis concerned influential factors in energy application have been considered during design of energy model in this article. The collected abundant data and economic indicator models by computation model obtain best solution for energy economics. A closed-loop predictive system is formed based on timing algorithm, to make the predictive model able to provide accurate prediction in the light of varied data [18][22][23].

To sum up above, taking the three types of energy sources as examples, in considering the platform needs to predict, statistic and analysis of the data, the overall model of energy is designed as shown in fig 3. The proposed computing models have been designed to target the following requirements of the Distributive Generation System (DGS) [2][24][25]:

- Analyze the situation and decide the data collection strategy and methodology on new and renewable sources. Collect and collate the relevant data required for modeling.
- Apply conceptual modeling for the design of integrated system like input on energy sources for the design of hybrid power plant to exploit maximum renewable energy sources at reasonable price.
- Either apply proposed models or in addition develop mathematical models for simulating environmental impact.
- Generate different scenarios ultimately to arrive at effective environment management plan with a view to support the decision makers.

The Control Design and Simulation Module (CDSM) provides a numerical simulation environment that enables users to test the model. CDMS is used to analyse the interactions between hybrid power solution comprises of mechanical-electrical systems [25] . Furthermore, the quality of existing models can be improved and other control strategies can be investigated by simulating deep-bar induction generators and more complex models of drive trains [15].

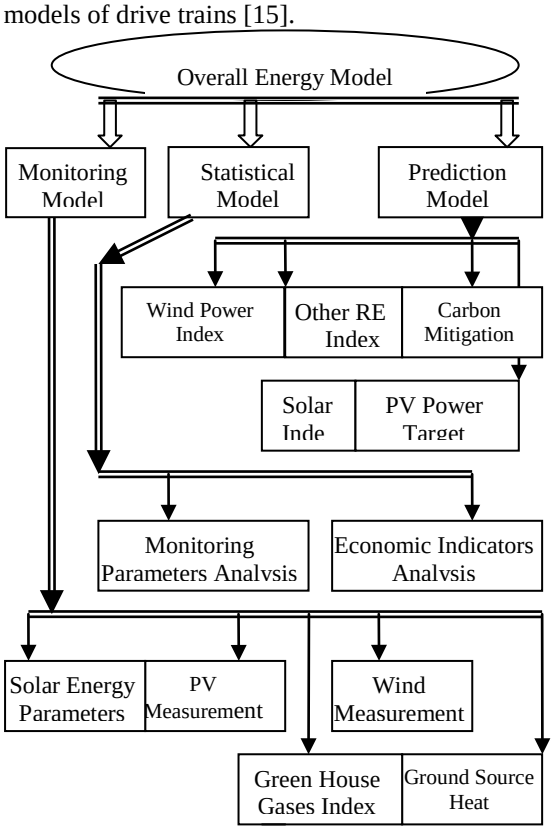


Figure 4: Integrated Energy Model

5. CONCLUSION

The computing proposes to develop algorithmic formulas for diversified renewable energy sources and building integrated projects. The proposed platform will also be able to conduct predictive analysis on the vast accumulated historical data, to aid finalization of the energy resource that is most economically and efficient. Furthermore, a statistical and analytical function is envisaged for this platform which can make comparative display of the same indicators of different projects or different indicators of the same project, hence providing a basis for popularization of renewable energy saving in different areas.

REFERENCES

[1]. IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation, IPCC, 2011

[2]. Aurobi Das, Dr. V. Balakrishnan "Energy Service Companies (ESCOs) to optimize Power in Peak Demand

Period in Hybrid Energy System- An Impact on Climate Change" 978-1-4244-5275-0/10/$26.00 ©2010 IEEE

[3]. Deepali Kamthania, G. N. Tiwari,"Determination of Efficiency of Hybrid Photovoltaic Thermal Air Collectors Using Aritificial Neural Network Approach for Different PV Technology" BIJIT - BVICAM's International Journal of Information Technology, Issue 7: (January-July, 2012 Vol.4 No.1)

[4]. Godfrey Boyle, "Renewable Energy" book published by Oxford University Press, 2004, ISBN978-0-19-958651-6

[5]. Hanane Dagdougui, Riccardo Minciardi, Member, IEEE, Ahmed Ouammi, Michela Robba, and Roberto Sacile, "A Dynamic Decision Model for the Real-Time Control of Hybrid Renewable Energy Production Systems", IEEE Systems Journal, Vol. 4, No. 3, September 2010, page 323-338.

[6]. Jinxu Ding and Arun Somani, "A Long-term investment planning model for mixed energy infrastructure integrated with renewable energy", 978-1-4244-5275-0/10/2010/ IEEE

[7]. L. Suganthia, Anand A. Samuelb, "Energy models for demand forecasting—A review" Renewable and Sustainable Energy Reviews 16 (2012) 1223– 1240

[8]. MNRE : National Solar Mission Document www.mnre.gov.in

[9]. M.E.H. Benbouzid, D. Diallo, Y. Amirat, H. Mangel and A. Mamoune "Entice students to power engineering using renewable energies undergraduate projects: example of development and application of wind turbines prototyping software under Matlab/Simulink®" 1, Rev. Energ. Ren. Vol. 8 (2005) 123 – 135

[10]. M. Soliman O.P. Malik D.T. Westwick, "Multiple model multiple-input multiple-output predictive control for variable speed variable pitch wind energy conversion systems", IET Renewable Power Generation, IET Renew. Power Gener., 2011, Vol. 5, Iss. 2, pp. 124–136

[11]. Peng Chen, Jie Liu*, Chongchong Yu, Li Tan, "Design and Implementation of Renewable Energy and Building Integrated Data Analysis Platform", 978-1-4244-4813-5/10 @2010 IEEE

[12]. Professor Lajos Göőz "Optimized Integration of Renewable Energy", EXPRES 2011 • 3rd IEEE International Symposium on Exploitation of Renewable Energy Sources • March 11-12, *2011*, Subotica, Serbia

[13]. R. Banosa, F. Manzano-Agugliarob, F.G. Montoyab, C. Gila, A. Alcaydeb, J. Gomezc, "Optimization methods applied to renewable and sustainable energy: A review" Renewable and Sustainable Energy Reviews 15 (2011) 1753–1766

[14]. Rajesh Kumar and Arun Agarwala, "Energy Computing Models for Techno-Economic Feasibility" INDIACom-2012; ISSN 0973-7529; ISBN 978-93-80544-03-8

[15]. Rajesh Kumar and Arun Agarwala, "RET Diffusion Model for Techno-Economics feasibility, International Conference on "SOLARIS -2012 - Energy security Global Warming and Sustainable Climate" organized by IIT Delhi and BERS on 7-9th February, 2012

[16]. REN21 Renewables 2012, Global Status Report, published by Paris: REN21 Secretariat.

[17]. S.C.Kaushik "Policy & Measure of economic efficiency, energy security and environment protection", Journal of Scientific & Industrial Research, Vol.66, November *2007*, pp 928-934

[18]. S.Saravanan, S.Vidya and Dr.S.Thangavel, "Design and Development of Multiple-Input Converter for Renewable Energy Integration", 978-1-61284-764-1/11/*2011* IEEE

[19]. T.V. Ramachandraa,b,B.V. Shruthib, "Spatial mapping of renewable energy potential" Renewable and Sustainable Energy Reviews, 11 (2007) 1460–1480

[20]. Tony B Nguyen, Ning Lu, and Chunlian Jin, "Modeling Impacts of Climate Change Mitigation Technologies on Power Grids" 978-1-4577-1002-5/11/$26.00 ©2011 IEEE

[21]. Deepali Kamthania and G. N. Tiwari, "Determination of Efficiency of Hybrid Photovoltaic Thermal Air Collectors Using Artificial Neural Network Approach for Different PV Technology", BIJIT - BVICAM's International Journal of Information Technology, Issue 7: (January-July, 2012) Vol.4 No.1,: ISSN 0973-5658

[22]. U. Navon, I Zur, D. Weiner, "Simulation model for optimising energy allocation to hydro-electric and thermal plants in a mixed thermaVhydro-electric power system", IEE PROCEEDINGS, Vol. 135, Pt. C, No. 3, MAY 1988

[23]. Rashmi Jha and A.K. Saini, "Process Benchmarking Through Lean Six Sigma for ERP Sustainability in Small & Medium Enterprises" BIJIT - BVICAM's International Journal of Information Technology, Issue 6: (July-December, 2011) Vol.3 No.2; ISSN 0973-5658

[24]. S.K.Muttoo, Sushil Kumar, "Data Hiding in JPEG Images" BIJIT - BVICAM's International Journal of Information Technology, Issue 1: (January-July, 2009) Vol.1 No.1 ; ISSN 0973-5658

[25]. Y. M. Atwa, E. F. El-Saadany, M. M. A. Salama, and R. Seethapathy Optimal Renewable Resources Mix for Distribution System Energy Loss Minimization, IEEE Transactions On Power Systems, Vol. 25, No. 1, February 2010 page 360-370

[26]. http://www.lindo.com/index.php?option=com_content&view=article&id=28&Itemid=4

Descriptive Analysis of Enrollment Data and Adaptive Educational Hypermedia

Nidhi Chopra[1] and **Manohar Lal**[2]

Submitted in November 2012, Accepted in April 2013

Abstract - As the world around is going through a technological revolution with the dawn of digital age, educationists are in some ways compelled to rethink the existing education system and its components. With the tools and the techniques available, nowadays it's imperative to reconsider how they can be used to improve educational institutions and associated bodies. Opportunities for knowledge discovery in educational data have increased tremendously with digital revolution now as compared to the scenario in the past. Educational data is becoming increasingly rich as more and more educational systems are going online and collecting large amounts of data. In this paper a study of an enrollment dataset is presented.

Index Terms - Data Analytics, Educational Data Mining, Enrollment data, Adaptive Educational Hypermedia.

1. INTRODUCTION

In this new digital age, the world of education has also gone under a major transformation. The new technologies and gadgets available help not only enrich and enhance the existing education system but also offer new opportunities and modes which can take the process of learning beyond institutions and allow people to learn on their own time and own terms. These new advances in learning have played a big role in this age of knowledge enhancement via different means and are clearly a sign that there is a need to rethink how the technology potential can be tapped to improve our education system [1, 2]. As of now most of the changes can be seen in the way information is stored, retrieved, distributed or provided to the students such as educational technology, e-learning portals, Learning Management Systems used in distance learning, blended learning & so on. Another emerging and associated area is educational data mining (EDM) where storage, retrieval and analysis of educational data sets can be leveraged to revolutionize education systems [3].

People in all fields and disciplines are becoming more and more informed. They are learning to observe, collect and interpret data trends around them to make better and informed decisions [4, 5]. Analysis of educational data sets is required to understand needs of current society and then also cut down costs in the process [4, 6]. In order to clearly define the framework and the needs for revolutionizing the current education system data have to be analyzed as a first step. In this paper such a study & some solutions are described.

[1, 2] *School of Computer and Information Sciences, Indira Gandhi National Open University, Delhi – 110068, India*
E-mail: [1]nidhichopra@ignou.ac.in and [2]mlal@ignou.ac.in

2. EDUCATIONAL DATA MINING

In the last few years EDM has emerged as a field of its own. The EDM community website [7] defines EDM as follows: "EDM is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in." Data mining (DM), or Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [8]. It finds applications in the fields of artificial intelligence, numeric & combinatorial optimization, business, management, medicine, computer science, engineering etc. [9]. DM largely consists of analyzing available sets of data to interpret, isolate the trends and patterns present in the data i.e. converting raw data into information. The trends obtained can be called as prediction or recommendations [10]. These can be used by educators, educational software developers, teachers, parents or students. However, it is largely understood that EDM methods are often different from standard DM methods. This is because of the non-independence and multilevel hierarchy found in educational data. For the same reason, it is increasingly common to see the psychometrics models being used in EDM [11]. DM is a part of Data Analysis. The outcomes of data based research can be descriptive or actionable, this study includes both.

DM can be visualized as a confluence of multiple disciplines where the background knowledge pertaining to the area of study is processed using tools pertaining to other disciplines such as – information science, database technology, statistics, machine learning & other related fields. Here the 'Area of Study' would be 'Education'. The data can be collected from students' use of interactive learning environments, computer-supported collaborative learning, evaluation, assessment or administrative data (web logs, library usage) from schools and universities. There are various challenges in the field of education like understanding choice of major, appropriate evaluation schemes, student drop out, retention, student unrest and crime, assessment of institution and educationists' goals like quality, access, cost, social and cultural biases. Educational efficacy can be measured and predicted using DM methods [12]. DM is a field which has originated from databases and Artificial Intelligence [13]. Understanding the current trends of our education system could point out towards the underlying issues and help us device an effective plan to address them.

Figure 1 shows broad two possible dimensions of EDM research wherein utilizing the data from point of view of educators and also from those studying the management/administrative aspect of EDM is considered.

Figure 1: Two possible dimensions of Educational Data Mining research.

2.1 EDM for Educators

An example for such type of EDM is PSLC (Pittsburgh Science of Learning Centre) [14, 15].

2.2 EDM for Administrator and Managers

Education Administrators also use EDM to understand more management related factors such as demographics, enrollment etc. An example of EDM for Admin is the presented case study of enrollment data consisting of 3020 record obtained from SRD (Student Registration Division) of IGNOU (Indira Gandhi National Open University). The data files are in dbf format (figure 2) and can be imported in MS-Excel (figure 3) using FoxPro. In dbf format, one column is shown in an entire screen in figure 2 whereas figure 3 now has a more readable tabular representation.

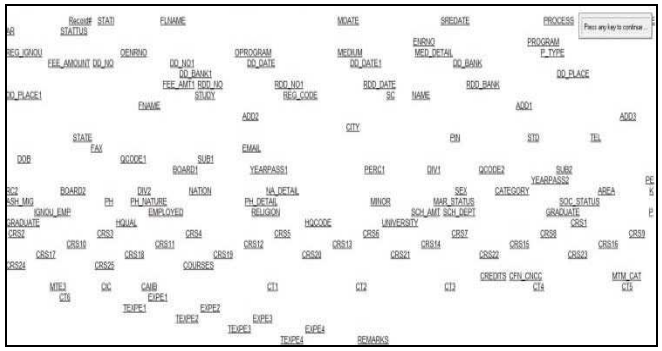
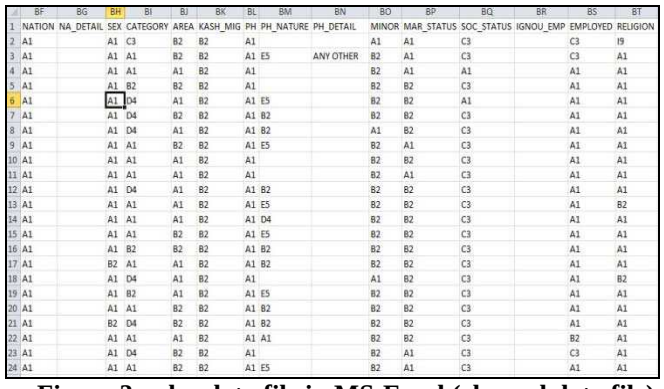


Figure 2: .dbf data file in MS-Visual Studio's Visual FoxPro (raw data file as received)



	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT
1	NATION	NA_DETAIL	SEX	CATEGORY	AREA	KASH_MIG	PH	PH_NATURE	PH_DETAIL	MINOR	MAR_STATUS	SOC_STATUS	IGNOU_EMP	EMPLOYED	RELIGION
2	A1		A1	C3	B2	B2	A1			A1	A1	C3		C3	I9
3	A1		A1	A1	B2	B2	A1	E5	ANY OTHER	B2	A1	C3		C3	A1
4	A1		A1	A1	A1	B2	A1			B2	A1	A1		A1	A1
5	A1		A1	B2	B2	B2	A1			B2	B2	C3		A1	A1
6	A1		A1	D4	A1	B2	A1	E5		B2	B2	A1		A1	A1
7	A1		A1	D4	B2	B2	A1	B2		B2	B2	C3		A1	A1
8	A1		A1	D4	A1	B2	A1	B2		A1	B2	C3		A1	A1
9	A1		A1	A1	B2	B2	A1	E5		B2	A1	C3		A1	A1
10	A1		A1	A1	A1	B2	A1			B2	B2	C3		A1	A1
11	A1		A1	A1	A1	B2	A1			B2	A1	C3		A1	A1
12	A1		A1	D4	A1	B2	A1	B2		B2	B2	C3		A1	A1
13	A1		A1	A1	A1	B2	A1	E5		B2	B2	C3		A1	B2
14	A1		A1	A1	A1	B2	A1	D4		B2	B2	C3		A1	A1
15	A1		A1	A1	B2	B2	A1	E5		B2	B2	C3		A1	A1
16	A1		A1	B2	B2	B2	A1	B2		B2	B2	C3		A1	A1
17	A1		B2	A1	A1	B2	A1	B2		B2	B2	C3		A1	A1
18	A1		A1	D4	A1	B2	A1			A1	B2	C3		A1	B2
19	A1		A1	B2	A1	B2	A1	E5		B2	B2	C3		A1	A1
20	A1		A1	A1	B2	B2	A1	B2		B2	B2	C3		A1	A1
21	A1		B2	D4	B2	B2	A1	B2		B2	B2	C3		A1	A1
22	A1		A1	A1	A1	B2	A1	A1		B2	B2	C3		B2	A1
23	A1		A1	D4	B2	B2	A1			B2	A1	C3		C3	A1
24	A1		A1	A1	B2	B2	A1	E5		B2	A1	C3		A1	A1

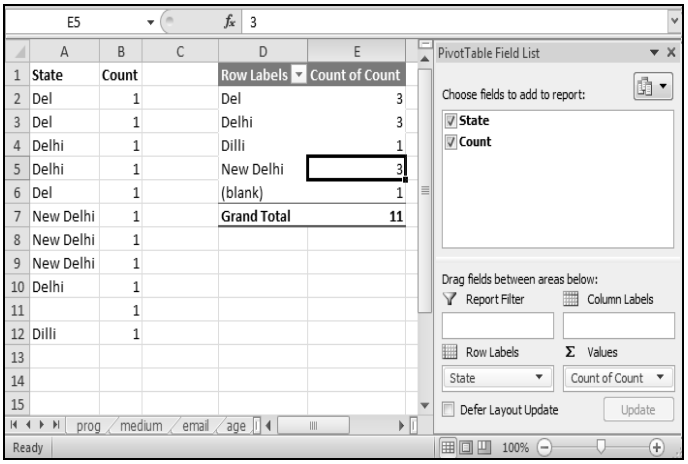
Figure 3: .xlsx data file in MS-Excel (cleaned data file)

For this research, enrollment data of disabled student for an entire year 2009 was obtained from them in January 2010. After data cleaning (using pivot table) some interesting patterns were obtained. The graphs obtained from this analysis are shown below and discussed in the next section. The research methodology has been followed from [16]. A wide variety of DM methods are available such as prediction, clustering, relationship mining, discovery with models, and distillation of data to obtain and present knowledge [17, 18]. Some relevant studies can be found in [19, 20 & 21].

3. STRATEGY FOR DATA CLEANING

Data parsing [22] is easy after importing file from FoxPro to MS-Excel because every column can be viewed separately now. The values are standardized already and discrete verifiable from university website & prospectus. Records were matched to see that there is no repetition of a student's enrollment number which is the primary key. Necessary transformation can be done e.g. to get age from date of birth. So, overall MS-Excel turned out to be a good tool for data cleaning.

'Cold start' is when a data miner has to start from scratch or 'zero' as in this study. Typical real world data sets which are unformatted (raw) need to go through data cleaning steps [22] to be successfully used in a study. After formatting the data appropriately, pivot table feature in excel (a statistical tool in MS-Office package) was used for Data Cleaning. Suppose the variable under consideration is 'State'. Various possible occurrences of State 'Delhi' can be counted as in figure 4. Blanks and wrong fields also got marked ('Del', 'Dilli'). This also explains that why sometimes local understanding of the database can be crucial.



	A	B	C	D	E
1	State	Count		Row Labels	Count of Count
2	Del	1		Del	3
3	Del	1		Delhi	3
4	Delhi	1		Dilli	1
5	Delhi	1		New Delhi	3
6	Del	1		(blank)	1
7	New Delhi	1		Grand Total	11
8	New Delhi	1			
9	New Delhi	1			
10	Delhi	1			
11		1			
12	Dilli	1			
13					
14					
15					

Figure4: Pivot table used to count variables& detect blanks and wrong code.

4. RESULTS

In this section some of the results are presented as obtained using pivot table feature of MS-EXCEL, while doing the data analysis conducted on disabled students of IGNOU who enrolled for various courses in the year 2009.

4.1 54.11% students are of young age group (figure 5).

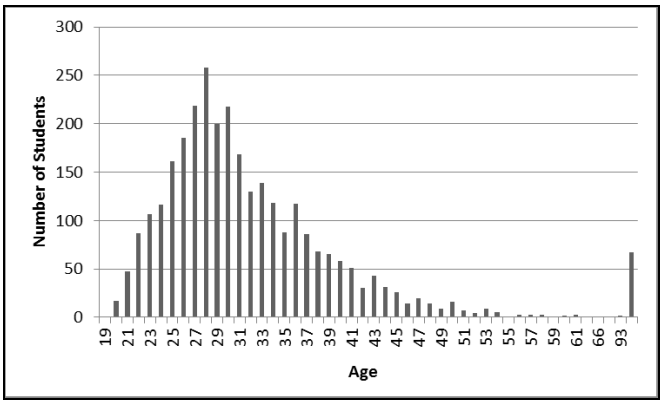


Figure 5: Graph showing age group distribution of students where last column represents 'Wrong code'.

4.2 37.11% students enrolled for Master of Political Sciences and paid an average fee (figure 6).

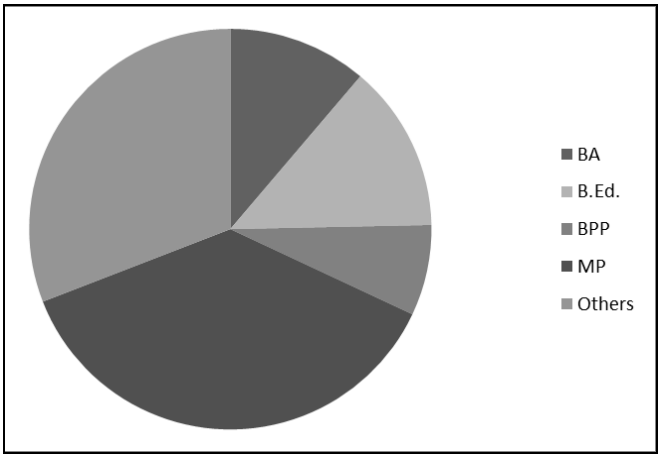


Figure 6: Distribution of courses/programs opted for by the students.

4.3 68. 01% opted for English Medium (figure 7).

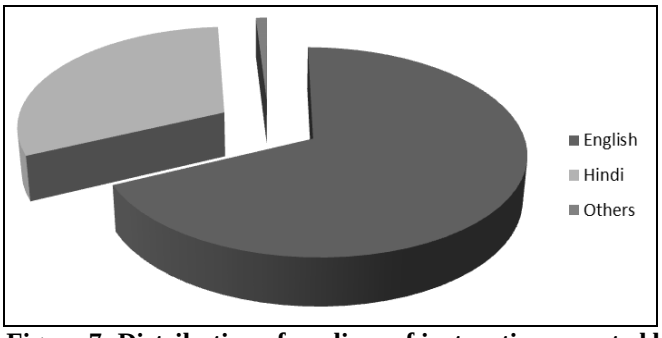


Figure 7: Distribution of medium of instruction as opted by students.

4.4 More than 70% students are male (figure 8).

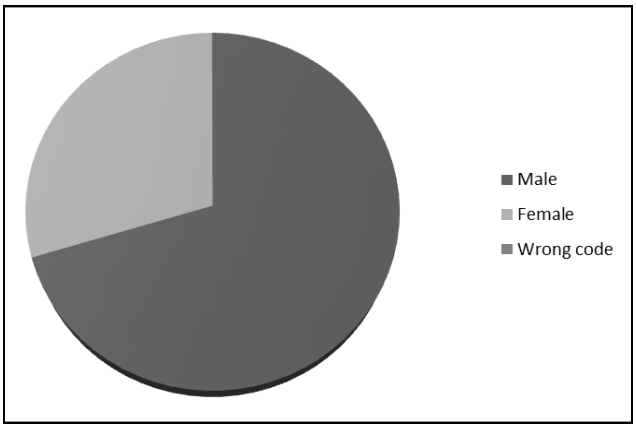


Figure 8: Pie chart showing distribution of gender across student population.

4.5 Most students had finished their previous educational qualification within the past decade as indicated in figure 9 below.

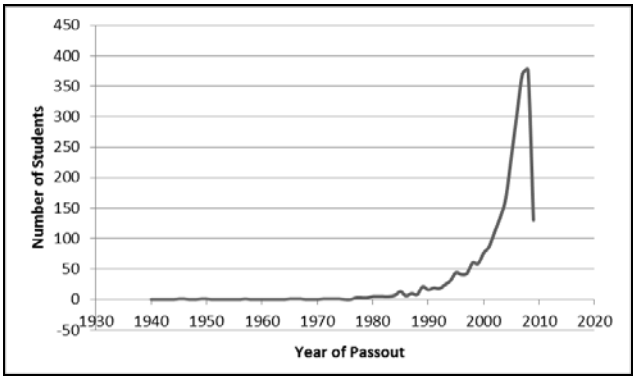


Figure 9: Distribution of students according to the year in which they finished their previous educational qualification.

4.6 54.4% of students are unemployed and 28.1% are employed by IGNOU itself. This shows that number of students who are pursuing education while being employed elsewhere is only 15.3% (figure 10).

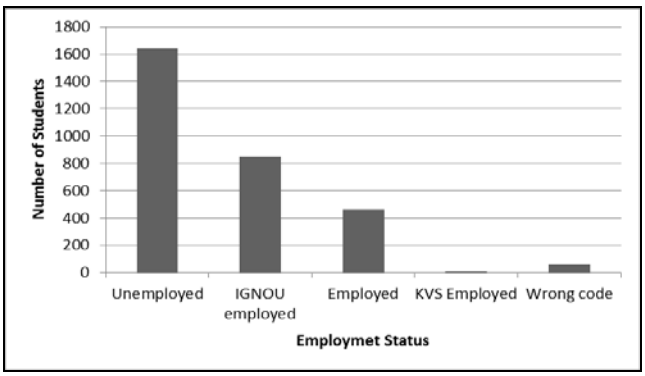


Figure10: Distribution of students accoring to employment status

4.7 Analysis of territory code of students address shows that 60% of students are from urban areas (figure 11).

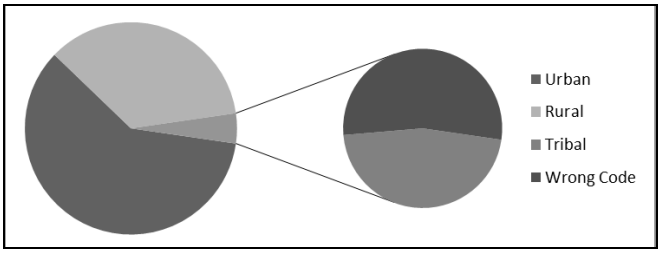


Figure 11: Distribution of student population as per territory code

4.8 To understand the accessibility, we analyzed the email ID field which showed that more than 71% students don't have email ids (figure 12).

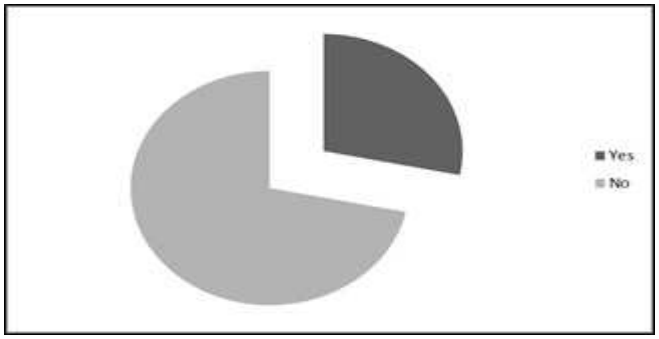


Figure 12: No e-mail ids indicate lack of access to internet and technology.

5. DISCUSSION WITH POSSIBLE ACTIONS

Above results indicate knowledge divide and digital divide. Better methods of increasing outreach are required.

5.1 Result 1 is self-explanatory. Figure 5 has an approximate shape of a normal distribution [23] as often exhibited in biological data [24]. This also resonates well with the model of our current education system where most people like to study or focus on their career value addition in their twenties or early thirties. This kurtosis curve is skewed to the left and a bit slant on the right.

5.2 Result 2 is due to the fact that these students find it easy to do humanities or social sciences courses because there is no help in the form of artificial limbs & training to use them in laboratories (sciences). Science laboratories have no accessibility equipment or area.

5.3 English medium books are comparatively easily available in India at higher education level. IGNOU however plans to launch courses in regional languages. More steps that can be taken are – to encourage translation of books/texts in all subjects and to make them accessible – brail translation, audio books (record and release), video field tours and online repositories of all these educational media.

5.4 Result 4 shows that disability is more common in males in this data sets. But raises more questions about infant & child mortality rate, gender biases or gender divide.

5.5 Result 5 is self-evident. Students are in their 20s, so they have mostly passed out in recent past. Figure 9 is shape of a chi-square distribution [23, 24], with one outlier - 'current' year pass outs.

5.6 Result 6 requires action from Governments to create accessible jobs to increase employment.

5.7 Result 7 indicates the possibility of urban area students having better accessibility to these courses i.e. Knowledge Divide. This needs to be verified by designing a focused future study for the same and improving awareness and accessibility in other areas as well.

5.8 No e-mail ids indicate lack of access to internet and technology for disabled students i.e. Digital Divide.

6. ACCESSIBILITY AND TRACKING

There is need to improve content delivery. It may help in decreasing digital and knowledge gaps. Currently quite a few e-learning and online information delivery platforms are designed with a "One-size-fits-all" approach. Existing distance education system lacks interactivity and can lead to lack of motivation and interest. There is a need for flexible education systems which can also provide guidance as per capacity & learning level [25].

Adaptive Educational Hypermedia (AEH) are flexible and customizable to provide appropriate lesson for each student. Here various views of the same material are created, as desired by the user This can be done by maintaining a student enrollment database combined with user behavior database using tools like link removal (figure 13), stretch text (figure 14) and course monitor (figure 15) for all those students of the university who are using online resources. These proposed tools utilize options set by user & also track and record actions of the user, which media type is chosen most often etc. Combined database form a user model [26] or a student model - goal, previous knowledge, previous performance, background, experience, preferences, stereotypes, user-supplied preferences supplied at run time, analysis of user actions & plan recognition or inference. Providing varying views of the same content is a paradigm shift away from "write once, use once" towards a middleware system [26].

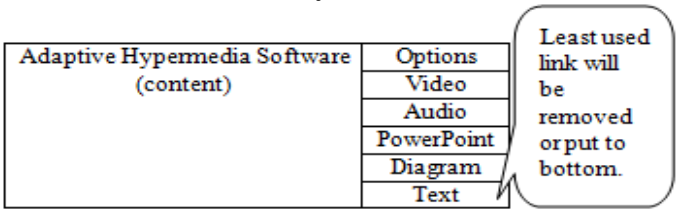


Figure 13: Link removal tool [27]

This link removal tool saves the time spent on looking for most preferred type of media by avoiding confusion. Same is the purpose of stretch text tool and it also makes the interface user friendly. Such tools adapt to the habits & needs of a user (HumanComputer Interface i.e. HCI).

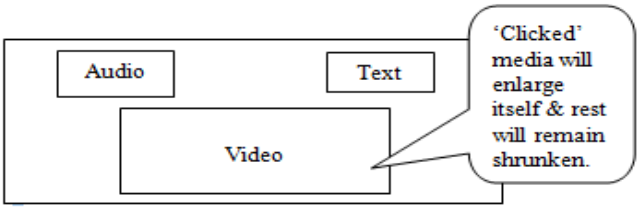


Figure 14: Stretch text tool [28]

Relations in the concept ensure that the student has a study guideline to follow and clearly knows the perquisites and the predecessors for each study. If certain specific perquisites are not fulfilled, the learner will be prompted for the same by course monitor tool. This tool is in accordance with Skinnerian or Linear Approach and can be combined with Programmed Learning. A rule for this can be as below. More components can be added – interest, repetition.

If c1.access = true then set c2.allow_access = true else c2.allow_access = false

If (c1.access = true or c1.test_passed = true) then set c2.allow_access = true else c2.allow_access = false (to include assessment & evaluation options)

Concepts from various disciplines can be combined – "many to many" approach.

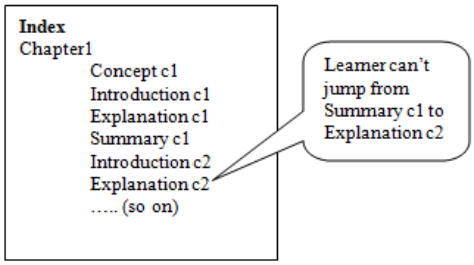


Figure 15: Course monitor tool [29]

7. SCALING IN EDM

A Data Mining problem can be solved through Generalization. To achieve a high degree and accuracy of generalization, a data miner needs a large number of records and more resources in terms of access, time, permissions, teams [30], better software, machines and related facilities as shown in figure 6. Over a longer period of time Agility in Data Analysis can be achieved through Software Re-Engineering. Real world implementations have high complexities [31]. Using tools like WEKA for data mining can give meaningless or useless results. In between steps are not shown as in a calculator so, a required level of understanding may not be obtained. Tools like SPSS require higher system configurations which may not be available to a researcher.

8. CONCLUSION AND FUTURE SCOPE

Analysis of educational data was discussed from the approach of administration. Understanding of the variables provided, exhibited digital divide and knowledge divide. AEH can be used to improve content delivery and may help in decreasing digital and knowledge gaps.It was observed that for developing EDM models, the data obtained should be focused, well organized to achieve effectiveness. At IGNOU where the dataset was collected from an administrative focus and without a preexisting problem statement, more data collection & further studies based on them are required to predict the trends. Studying educational data sets can aid in suggesting pedagogies (teaching methods), site modification, intelligence services & page recommendations in the long run [22, 32].

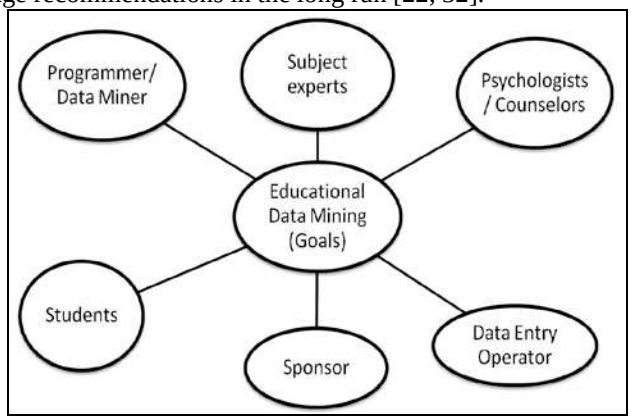


Figure 16: Team work in EDM

To improve the analysis and usability of enrollment data, the enrollment form at universities should be improvised to collect relevant/targeted data fields [33] such as degree of disability, monthly/annual income of the family and other personal data of family, individual/student and assessment. Online enrollment can facilitate collection of data for analysis. Such e-forms can be made adaptive in nature. Background information and performance of every pupil can be assessed throughout the academics and the employment/career to clearly identify any patterns and correlations in the data.

Another possible study can be to analyze the assessment data [34, 35] & log files of these students to find non-independence and multilevel hierarchies in educational data [36]. Such analysis can help us provide more useful insights in the education system of IGNOU for disabled students and to understand the factors affecting students' learning and career path development.

ACKNOWLEDGEMENTS

Authors thank SRD, IGNOU for providing disabled students' enrollment data of the year 2009,for this study.

REFERENCES

[1]. A. Collins and R. Halverson - Rethinking education in the age of technology: the digital revolution and schooling in America. Teachers College Press, 2009.

[2]. C. Romero & S. Ventura. "Educational data mining: A survey from 1995 to 2005". Expert Systems with Applications, 2007, 33(1), 135-146.

[3]. F. Siraj, & M. A. Abdoulha. "Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining". Third Asia International Conference on Modeling & Simulation, 2009. 413-418. IEEE.

[4]. S. Khan and S. Kumar – "Optimization of Material Procurement Plan – A Database Oriented Decision

Support System", BIJIT - BVICAM's International Journal of Information Technology, 2012.
[5]. Q. A. Al-Radaideh, E. M. Al-Shawakfa & M. I. Al-Najjar.. "Mining student data using decision trees." International Arab Conference on Information Technology, 2006, Yarmouk University, Jordan.
[6]. G. Peters, R. Tagg, and R. Weber. "An application of rough set concepts to workflow management." Proceedings of the 3rd international conference on Rough sets and knowledge technology, 2008, 715-722.
[7]. www.educationaldatamining.org(accessed on 19 December 2011).
[8]. I. H. Witten and E. Frank - Data Mining – Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2005.
[9]. S. V. Chande and M. Sinha – "Genetic Algorithm: A Versatile Optimization Tool", BIJIT - BVICAM's International Journal of Information Technology, 2008.
[10]. C. Vialardi, J. Bravo, L. Shafti, & A. Ortigosa, "Recommendation in higher education using data mining techniques." Proceedings of the 2nd International Conference on Educational Data Mining, 2009, 190-199
[11]. R. S. J. D. Baker, K. Yacef, K. "The state of educational data mining in 2009: A review and future visions." Journal of Educational Data Mining, 2009, 1(1), 3-17.
[12]. T. M. Mitchell - Machine Learning. McGraw-Hill. 1997.
[13]. T. Munakata - Fundamentals of the New Artificial Intelligence, Neural, Evolutionary, Fuzzy and More, Springer, 2008.
[14]. http://learnlab.org/(accessed on 17 August, 2011).
[15]. http://learnlab.org/opportunities/summer/ presentations/2011/PSLCSummerSchoolPostersAndFirehoseSlides2011.zip(accessed on 17 August, 2011).
[16]. F. C. Dane - Research Methods, Brooks/Cole Publishing Company, 1990.
[17]. R. S. J. D. Baker, Data mining for education. International Encyclopedia of Education, 2010, 7, 112-118.
[18]. J. Han and M. Kamber - Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2006.
[19]. N. Delavari, M. R. A. Shirazi & M. R. Beikzadeh, "A new model for using data mining technology in higher educational systems." Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training, 2004, 319-324 IEEE.
[20]. N. Delavari, M. R. Beikzadeh, & S. Phon-Amnuaisuk, "Application of enhanced analysis model for data mining processes in higher educational system".6th International conference on Information Technology Based Higher Education and Training, 2005, F4B-1. IEEE.
[21]. C. Vialardi, J. Chue, A. Barrientos, D. Victoria, J. Estrella, A. Ortigosa, J.Peche, "A case study: data mining applied to student enrollment". Proceedings of Third Educational Data Mining Conference, Pennsylvania, USA 2010, 333-335
[22]. M. A. Butt, M. Zaman, Data Quality Tools for Data Warehousing: Enterprise Case Study, IOSR Journal of Engineering, 2013, 3(1), 75-76.
[23]. S. Lipschutz and J. Schiller - Introduction to Probability and Statistics, Tata McGraw Hill, 2005.
[24]. P.S.S.S. Rao and J. Richard - An Introduction to Biostatistics- A Manual for students in Health Sciences, 3rd Edition, PHI Pvt. Ltd. 1999.
[25]. Y. Liao, The Application of Web Mining in Distance Education Problem, Proceedings of the 2nd International Symposium on Computer, Communication, Control and Automation, China 2013, Atlantis Press.
[26]. J. Eklund, R. Zeiliger, Navigating the Web: Possibilities and Practicalities for Adaptive Navigational Support. Proceedings of Ausweb'96: The Second Australian World Wide Web Conference. 1996, Southern Cross University Press.
[27]. A. I. Cristea, C. Stewart, Automatic authoring of adaptive educational hypermedia. Web-Based Intelligent e-Learning Systems: Technologies and Applications, 2005,IDEA Publishing group, Zongmin Ma.
[28]. D. Paul, P. Brusilovsky, and G.-J. Houben.Adaptive hypermedia: from systems to framework. ACM Computing Surveys (CSUR) 31.4es , 1999: 12.
[29]. A. I. Cristea, A. D. Mooij. "Designer adaptation in adaptive hypermedia authoring."Proceedings of International conference on Information Technology: Coding and Computing IEEE, 2003.
[30]. A. Cockbum - Agile Software Development, Pearson Education Inc., 2002.
[31]. N. Aggarwal, N. Prakash and S. Sofat – "Mining Techniques for Integrated Multimedia Repositories" BIJIT - BVICAM's International Journal of Information Technology, 2008.
[32]. A. S. Alvi, M. S. Ali., "Revival of Tutor Model: A Domain Independent Intelligent Tutoring System" (ITS)", BIJIT - BVICAM's International Journal of Information Technology, 2010 Vol.2 No.1
[33]. R. Srivastava, I. Srivastava, "Computer and Internet use Among Families: A Case of Botswana", BIJIT - BVICAM's International Journal of Information Technology,2009 Vol.1 No.1
[34]. A. Merceron, K. Yacef, Educational data mining: a case study. Artificial Intelligence in education: supporting learning through Socially Informed Technology.–IOS Press, 2005, 467-474.
[35]. F. Castro, A. Vellido, A. Nebot, F. Mugica, F. Applying data mining techniques to e-learning problems. Evolution of teaching and learning paradigms in intelligent environment, Springer Berlin Heidelberg, 2007, 183-221.
[36]. J. Luan, Data mining and its applications in higher education. New directions for institutional research, 2002, vol. 2002 No. 113, 17-36.

Knowledge Representation in *pAninI* Framework Using Neural Network Model

Smita Selot[1], Neeta Trpathi[2] and **A.S Zadgaonkar[3]**

Submitted in October 2012; Accepted in February 2013

Abstract - Knowledge representation is base for expressing semantic content of input in intelligent information retrieval systems. Identification of semantic requires processing of input language at various levels. To make system understand text or speech is a challenging task as it involves extracting semantics of the language which itself is a complex problem. At the same time languages posses with multiple ambiguities and uncertainty which needs to be resolved at various phases of language processing. Level of understandability depends upon the grammar, syntactic and semantic representation of the language and methods employed for these analysis. Processing depends on the type of language, grammar of the language, ambiguities present and size of corpus available. Order free language posses different features as compared to rigid order language. Most of the Indian languages are order free; hence mechanism for such language needs to be formulated. One of the ancient Indian Sanskrit grammarians, pAninI has defined grammar of Sanskrit language in such a way that it is suitable for computational analysis. Six main semantic class identified under this theory is a baseline model for knowledge representation. This paper exploits the features of the language, applicability of rules and resolving ambiguities using neural network model. A hybrid model incorporating the features of rules based and neural network the is designed and implemented for pAninI based semantic analysis, generating case frames as output.

Index Terms - pAninI Grammar framework, Knowledge Representation, Case Frame, Natural Language Processing, Semantic.

1. INTRODUCTION

Knowledge representation is a technique to represent the meaningful and logical content embedded in the language; in a structured form. Development of such tool requires an exhaustive analysis of input language at syntactic and semantic level with capacity to handle ambiguities at each level. Natural languages are not so natural for computer processing; hence a KR tool acts as bridge between the natural language and understanding of language by machine. Development of such tool is heavily guided by language processing techniques and type of language. Order free language posses different characteristics than rigid order language. As most of the Indian languages are order free, they require different mechanism to handle their processing. KR, Natural Language Processing (NLP) and Information Retrieval (IR) are close module of such applications as depicted in Figure 1.

[1]SSCET, Bhlai, [2]SSITM, Bhilai, India
[3]C V Raman University, Bilaspur, *E-mail:* [1]*sselot@sify.com*

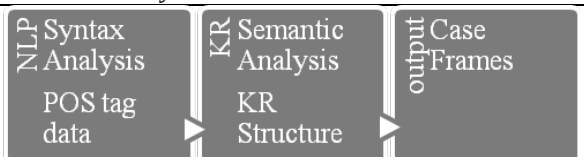


Figure 1: Inter relation between NLP and KR

Statistical methods are applied for syntactic analysis of Indian language with Hidden Markov Model(HMM) [12], support Vector machine (SVM) being popular statistical Technique [4] [19]. Application of Neural Network for classification task is less observed as both are complex domain. This paper presents a method for generation of Case Frames (CF) as KR structure for Sanskrit Language under *pAninI* framework. Method identifies semantic role of each word with respect to action or verb present in the sentence, there by presenting a verb-argument relation. Six main semantic classes are defined under *pAninI* framework. Identification and classification of word into one of the class is achieved by analyzing suffix attached to word. Identified class along with word is stored in KR structure called CF. However while performing the classification one suffix may map into multiple domain resulting into conflicting output. Such conflict is resolved by training neural network for ambiguous cases. Non conflicting cases are handled by one-to-one *vibhakti_kArka* mapping resulting into a hybrid model for case frame generation. This paper describes the concept of *pAninI* grammar for semantic analysis, database of suffix, algorithm and solutions for conflict cases. KR based system are widely used in applications like translation system, learning algorithm and question answer based system

2. *pAninI* GRAMMAR

One of the ancient languages of the world, Sanskrit, has well defined grammatical and morphological structure which precisely defines the relation of suffix-affix of the word with the syntactic and semantic classification of the sentence [2][3] [11]). Such analysis leads to development of KR structure. For order free language like Sanskrit, processing is quite interesting as suffix based analysis reveals syntacto-semantic features of the sentence. Sanskrit is analyzed from computational perspective on vedic text [7] as well as capability of *pAninI* grammar is equivalent to finite state machine [8]. Development of automatic segmentiser is an effort in this field [13]. Hindi and Arabic clauses are also analysed from *pAninian* aspect [14]. Parallelism of *pAninI* in field of computer science is well explained [15]. Rule based POS tagger developed at JNU,Delhi uses lexicon and displays all possible outcome for conflicting cases [9]. This paper explains processing of Sanskrit for classifying words in one of six semantic roles defined by *pAnini* under *kAraka* theory implementing a novel approach –Neural Network.

Generally, dictionary of words is maintained and each word is mapped to find its respective syntactic category. As *pAninI* has identified the syntacto-semantic information of the word by the suffix attached to the word, instead of maintaining dictionary of words, lexicon of suffix is sufficient for extracting features. *kAraka* roles are similar to case based semantics required for event-driven situations, where entities like agent , object, location are identified with respect to each event [6] [10].

pAninI, an ancient Sanskrit grammarian has given nearly 4000 rules called *sutra* to describe behavior of the language in the book called *asthadhyAyi;* meaning eight chapters [10]. Ancient old *kAraka* theory rules are in parallel with finite state machine [8] and concept is being extended for English language [20]. It describes transformational grammar which applies sequence of rules to transform root word to number of dictionary words. From small set of root words, millions of words are generated by firing set of rules. For highly inflectional language like Sanskrit, sequence of declension tables are memorized in such a way that similar ending words follow the same declension. Hence, if one table is memorized, number of words can be generated if their base word falls under same group. This structural representation in optimum form is used to identify the semantic class of the word. In Sanskrit language, fundamental six roles, given by *pAninI* as *kAraka* values, are key semantic component of a sentence as described in Table 1.

SN	Case	*kAraka*	*Vibhakti*	Meaning
1	Nominative	*Karta*	*prathamA*	Agent
2	Accusative	*Karma*	*dviteeyA*	Object
3	Instrumental	*karNa*	*triteeyA*	Instrument
4	Dative	*sampradAn*	*Chaturthi*	Recipient
5	Ablative	*apAdAn*	*Panchai*	Departure
6	Locative	*adhikaraN*	*Saptami*	Place

Table 1: Six *kAraka* in *pAnanian* model.

Suffix driven analysis is performed by mapping the suffix to database which contains suffix and a key number. Key is designed in such a way that it contains all the syntactic information as per grammar of the language

3. KEY NUMBER DESIGN FOR SUFFIX

All the nouns in the language follow nominal declension tables for each category of word. For example all '*a*' ending word follow the declension given in Table 2 with word as '*rAma'* and Table 3 shows the suffix attached to the word.

Vibhakti	***Ekvachan***	***Dwivacahn***	***Bahuvachan***
1	*rAmH*	*rAmau*	*rAmAH*
2	*rAmam*	*rAmau*	*rAmAn*
3	*rAmen*	*rAmAbhyAm*	*rAmaiH*
4	*rAmAya*	*rAmAbhyAm*	*rAmebhyH*
5	*rAmAt*	*rAmAbhyAm*	*rAmebhyH*
6	*rAmasya*	*rAmayoH*	*rAmAnAm*
7	*rAmen*	*rAmayoH*	*rAmeShu*

Table 2: Declension set for *rAma*

Vibhakti	***Ekvachan***	***Dwivacahn***	***Bahuvachan***
1	*H*	*Au*	*AH*
2	*Am*	*Au*	*An*
3	*En*	*AbhyAm*	*aiH*
4	*Aya*	*AbhyAm*	*ebhyH*
5	*At*	*AbhyAm*	*ebhyH*
6	*Asya*	*yoH*	*AnAm*
7	*En*	*yoH*	*eShu*

Table 3: Suffix for all *a* ending word

Each row corresponds to a *vibhakti* value and column represents the number or *vachan*. Unlike English language, which contains only singular and plural, Sanskrit has singular as *ekvachan* , plural as *bahuvachan* and two in number is labeled as *dwivachan*. *Vibhakti* is related to *kArka* values, Suffixes present in first row or *vibhakti* is *kartA kAraka* (agent). Likewise each row represents a *kArka* role as given in the Table 3.1.Sixth *vibhakti* is not included in Table 3.1 as it is *sambandh kAraka* which has relation with its immediate argument and not related to verb directly, hence not considered as *kAraka* by *pAninI*.

Four digit key number schemes for noun suffix is designed as given in Fig 2

x ending	Gender	Vibhakti	Number

Figure 2: Four digit number scheme for noun

Type of ending is in first column where 9 different type of ending is considered with values in range 1-9.Gender are masculine, feminine and neuter with values 1,2 and 3. Seven *vibhakti* from 1 to 7 and three number from 1to 3 are considered. For example suffix *'am'* is present in second row, first column as given in Table 2.2; it is assigned the value 1121 where description of each digit is as follows:

1: *a* ending
1: Masculine gender
2: *dvitIyA vibhakti*
1: *ekvachan*

On similar guideline, five digit number schemes is designed for storing verb suffix [16]. In Sanskrit grammar, verbs are classified into ten groups called *gan* represented by most significant place in the number scheme. When a root word joins with the suffix (*pratya*), some changes takes place at the junction. With respect to these changes verbs are classified into nine different *gan.* Second digit from left denotes *pad* which occur in three different forms- *Atmnepad*, *parsmaipad* and

ubhaypad. Verbs whose outcome is for another person, they fall under
Parasmaipad and verbs whose outcome is for one self come under *Atmnepad.* Verbal words whose outcome is for both, other person and one self, they come under *ubhaypad*. Time in which action takes place is given in various tenses. There are 10 different tenses in Sanskrit [5]. Giving the context of the person is *purush* and number is *vachan.* A five digit number scheme for verb is presented in Fig 3.

Gan	*Pad*	Tense and mood	Person	Number

Figure 3: Five Digit Number Scheme for Verb Suffix

Verbs in Sanskrit decline with respect to *gan*, *pad*, tense, person and number. These are influential parameters as they govern the behavior of the nouns in the sentence. Range and example values associated with each field are described in Table 4.

Digit	*Gan*	*Pad*	Tense & mood	Person	Number
Range	0-9	1-3	0-9	1-3	1-3
Example	*bhavAdigan,*	*Parasmaipad*	*Latlakar*	*pratham puruSh*	*Ekvachan*

Table 4: Range and Example Values for Each Digit Position of Verb Number Scheme

For example *paThati* (to read), has suffix *ti* which extracts the number 11011 from database there by giving the following information:

1-*bhavAdigan*;
1-*parasmaipad*;0-*latlakAr*(Present tense);
1-*pratham puruSh* (Third Person);
1-*ekvachan* (singular) .

Pronouns decline in manner similar to noun and total number of pronoun is less, hence set of most commonly used pronoun are stored in a separate database with their key values[17]. Number scheme for pronouns is given in Fig 4.

P_number	Gender	Vibhakti	Number

Figure 4: Four Digit Number Scheme for Pronoun

P_number identifies a particular pronoun like *serva, sH* etc. For example, 1 is given to *tad*, meaning 'that' in English. Rests of digit have same values as for noun. All the pronouns stated in *rachanAnuvAdakaumudI* are considered with nearly 400 entries in database [5]. Some words which do not change their form under any condition, they are termed as *avyay.* List of these words are maintained separately.

4. ALGORITHM FOR CASE FRAME

Objective is to generate the CF by identifying the semantic role (*kAraka* value) of each word with respect to action in a given sentence. Every word within the sentence is searched in *avyay* list. On unsuccessful search in the list, word is mapped in pronoun, verb suffix and then noun suffix database. This order has been followed as the quantity of words in each category follows an ascending order.

After check in *avyay* list; word w_i, is mapped in pronoun database; If found, its semantic role is identified by extracting the 4th digit from its key and performing *vibhakti kAraka* mapping on it. Otherwise next look up is performed on verb database (VDB) followed by noun database (NDB).

String is processed in reverse order from right to left. Last character of the word w_i is identified (x) and all the suffix which has x as their last character is extracted from the VDB and stored it in a set. If any one of the value from this set matches as suffix in word; word w_i, it is added to list of mapped suffix. If this list contains one element then a unique mapping has been found, else multiple matches are discovered. In case of multiple matches, splitter algorithm is activated to check for the category of word. Splitter breaks the word in base word and suffix. If base word is present in lexicon of verbal base, current word is tagged as action entity. If no match is found in verbal base then check for noun is performed.

A similar mapping process is also performed for nouns, but this mapping process face a problem as occurrence of suffix in database is not unique; at time multiple matches are obtained. It is due to intergroup and intragroup redundancy of suffix. Occurrence of same suffix within one declension table is intragroup redundancy and occurrence of same suffixes across tables is intergroup redundancy. Depending upon frequency of occurrence and redundancies, suffixes are divided into three classes. Class I identifies unique occurrence of suffix, class II identifies intergroup redundancy and class III identifies intragroup redundancy.

Class 1: Unique suffix
Suffix with frequency of occurrence =1.
Format of the data is
(<key number>, <suffix>, <frequency of occurrence>)
Example: (1111, 'H', 1)

Class 2: Intragroup redundancy
Suffix with frequency of occurrence greater than 1 and same *kAraka* value
Example: (1131, 'en', 2)(1331, 'en', 2), suffix 'en' have same *kAraka* value 3.

Class 3: Intergroup redundancy
Suffix with frequency of occurrence greater than 1 and different *kAraka* value.
Example: (1112, 'au', 4) (1122, 'au',4) (2171, 'au', 4)(3171, 'au', 4), suffix 'au' has *kAraka* values 1, 2, 7, 7.

kAraka value is the 3rd digit in the key from left. Categorization of the suffixes is presented in Fig 5

All x-ending suffixes s_i, are extracted from NDB and stored in a set. If suffix belong to class I, it is easily given a *kAraka* role. If more than one suffix is present, then for every suffix s_i, word w_i is split in base word and suffix using splitter algorithm. This

base word is searched in lexicon of base words. If a match is found, then s_i and its key number are stored as final suffix and final number in a list. Class II type Intergroup redundancy of the suffixes are handled by splitter routine. Ambiguity related to class III may still exist. To resolve such cases conflict resolution using neural network is applied. *Vibakti kAraka* mapping is applied to 4th digit of this number and a semantic tag is assigned to the word. All the words with their semantic tag are stored in the case frame.

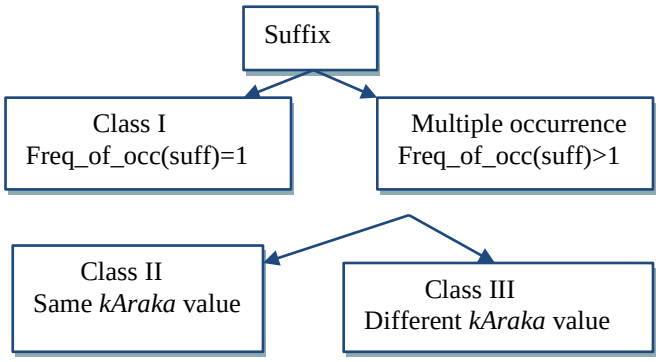


Figure 5: Categorization of Suffix with Respect to their *kAraka* Value

Out of the total 252 suffixes for noun, number of suffix belonging to each class is given in Table 5.

Quantitative parameter for noun suffix	Quantity
Number of suffix	252
Number Unique suffix	120
Class 1	55
Class 2	22
Class 3	43

Table 5: Class wise Quantification of Suffix Data

Algorithm for hybrid model

vlist = lexicon of verb bases
nlist= lexicon on nominal bases
w_i = word under process
s_i = suffix
x= last character of word under process.
verb_xmatch: set of all suffix ending with *x*
verb_suffix: set of all suffix in verb_xmatch which map as suffix in word w_i
noun_xmatch: set of all suffix ending with *x*
noun_suffix: set of all suffix in noun_xmatch which map as suffix in word w_i
frame: is structure for storing elements like action, agent, object etc.

Pseducode

Input sentence S.
for each word w_i € S
match w_i in pronoun database
if match found ***then***
w_i.cat=pronoun
get key number from database of w_i
vibhakti kAraka mapping
Frame.<case>= w_i
Break
else
last char of w_i = x
verb_xmatch = s_i from VDB such that last_char(s_i) = x
verb_suffix= suffix in verb_xmatch that appear in word w_i as suffix
if verb_suffix != NULL ***then***
for each s_i € verb_suffix ***do***
(base, s_i)=splitter(w_i s_i)
If base € vlist ***then*** frame.action = w_i
end for
endif
if match not found in VDB, ***then*** check in NDB
noun_xmatch = all suffixes s_i from NDB such that last_char(s_i)=x
noun_suffix = all suffix in noun_xmatch that appear in word w_i as suffix
num_set = respective number of matched suffix
if noun_suffix != NULL ***then***
for each s_i € noun_suffix and num_i € num_set
(base, suffix)=splitter(w_i $s_{i,}$, $num_{i,}$)
if base € nlist ***then***
Identify the vibhakti of the word
If class =1 ***then*** one value of *vibhakti* is obtained ***else***
If class=2 ***then*** one value of *vibhakti* is obtained ***else***
If class=3 ***then*** more than one value of *vibhakti* is obtained
Call for conflict resolution using NN
Perform *vibhakti-kAraka* mapping
Store the *kAraka* value as semantic role of the word
frame.<case> = w_i
end for
endif
Return frame

Results of the algorithm is discussed in last section, here conflict cases under class II are resolved using neural network. , discussed in next section.

5. CONFLICT RESOLUTION USING NN

For nouns in Sanskrit, intragroup redundancy of type III can be resolved using either statistical methods or NN based method. Statistical techniques require large corpus of data, due to lack of large size data with good vocabulary coverage, NN is implemented for conflict resolution. Back propagation algorithm with three layers is used to train the system for conflict cases. A set of pre annotated text is prepared which contain the suffix, category and *vibhakti* for each word of sentence. Sample of the annotated text is presented in Fig 6.

tvam/am.p.1 bhojanam/am.n.2 pachasi/si.v.x.pach|
devH/H.n.1 vanam/am.n.2 gachchhati/ti.v.x.gachch|
hariH/H.n.1 putrAya/Aya.n.4 bhojanahm/am.n.2
pachati/ti.v.x.pac|
rathavAhakH/H.n.1 aShvebhyH/ebhyH.n.4 ghAsam/am.n.2
anayati/ti.v.x.anaya|
idam/am.p.1 chAtrasya/asya.n.6 pustakam/am.n.2 asti/ti.v.x.as|
devH/H shakten/en.n.7 gRAmam/am.n.2

Figure 6: Sample annotated data

Most common ambiguous cases for *vibhakti* or *kAraka* value fall under four main domain (1,2),(1,2,7),(3,4,5) and(4,5). NN takes features of corpus as input and final *vibhakti* or *kArka* value as output. Features selected for training network is given in Table 6

Parameter	**Feature type**
Candidate suffix	Morphological feature
Candidate word category	Syntactic feature
Verb suffix	Morphological feature
Verb prefix	Morphological feature
Verb root	Lexical feature
Successive word	Context based feature
Previous word	Context based feature
Probability vector for suffix	Corpus based feature

Table 6: Feature selected for NN training

A NN based system takes the input in numerical form; hence the word features are converted into suitable numerical value. Mapping of features into numerical values is shown in Table 7.
Input coding algorithm reads the pre annotated text and generates the data for training the neural network.
BPN is feed forward multi layer network consisting of mainly three layers. Algorithm uses two passes - forward and backward pass. In the forward pass, inputs are multiplied by respective weight and a bias added to it. Weighted sum of input along with bias is fed as input to hidden layer. Hidden layer uses a squashing function to limit the output value in desired range. Output from hidden layer is multiplied by respective weight and fed to output layer. Sigmoid function is used to limit the range of output. Output obtained is compared with actual value and error is calculated as difference of the two. Error is a measure of difference between actual and the desired output. This calculated error is propagated back in the backward pass. To improve the performance of the network, weights are modified as a function of propagated error. In the forward pass, weights of the directed links remain unchanged at each processing unit of the hidden layer. For n input values weighted sum is obtained and sigmod function is applied to this weighted sum. Time taken to train the network is directly proportional to size of data. If number of neurons is increased, training time increases. Classification of *pAninI kAraka* with NN require large size corpus for training. Hybrid model overcomes the problem of large training time by classifying the word with their *vibhakti* value in non-conflicting situations and applying NN under conflicting situations only. This requires a small set of data and network is trained for conflicting classes only, thereby reducing the time. All the cases under same conflicting domain require same network. Major conflicting domains are (1, 2) (3, 4, 5) (6, 7).As data set for each conflicting case focuses on limited set of suffix, small data size is sufficient. For example *am* ending suffix fall in two class 1 and 2; NN was trained on these cases and result of the cases is presented next section.

Category	
1	Pronoun
2	Noun
3	Verb
4	Avyaya
5	Adverb

Prefix(verb)	
1	up
2	anu
3	adhi
4	a
5	abhini

Verb root			
1	gachch	11	dhuh
2	pach	12	yAch
3	kar	13	danD
4	bhu	14	rudh
5	anay	15	pricchch
6	As	16	Chi
7	pashy	17	bru
8	vas	18	Ji
9	dly	19	math
10	yachch	20	suSh

Vibhakti-karka	
1	Agent
2	Object
3	Inst
4	Cause
5	From
6	Relation
7	Location

Suffix(verb)	
1	ti
2	si
3	at
4	Ami
5	anti
6	AmH
7	te
8	tH

Suffix(noun)	
0	No suffix
1	Am
2	H
3	Aya
4	ebhyH
5	Asya
6	en
7	e

Table 7: Sample set of encoded values

6. CONFLICT RESOLUTION USING NN

For nouns in Sanskrit, intragroup redundancy of type III can be resolved using either statistical methods or NN based method. Statistical techniques require large corpus of data, due to lack of large size data with good vocabulary coverage, NN is implemented for conflict resolution. Back propagation algorithm with three layers is used to train the system for conflict cases. A set of pre annotated text is prepared which contain the suffix, category and *vibhakti* for each word of sentence. Sample of the annotated text is presented in Fig 7.

tvam/am.p.1 bhojanam/am.n.2 pachasi/si.v.x.pach|
devH/H.n.1 vanam/am.n.2 gachchhati/ti.v.x.gachch|
hariH/H.n.1 putrAya/Aya.n.4 bhojanahm/am.n.2 pachati/ti.v.x.pac|
rathavAhakH/H.n.1 aShvebhyH/ebhyH.n.4 ghAsam/am.n.2
anayati/ti.v.x.anaya|
idam/am.p.1 chAtrasya/asya.n.6 pustakam/am.n.2 asti/ti.v.x.as|
devH/H shakten/en.n.7 gRAmam/am.n.2 gachchhati/ti.v.x.gachch|
sH/H.p.1 mitre/e.n.7 vishvAsam/am.n.2

Figure 7: Sample annotated data

Most common ambiguous cases for *vibhakti* or *kAraka* value fall under four main domain (1,2),(1,2,7),(3,4,5) and(4,5). NN takes features of corpus as input and final *vibhakti* or *kArka* value as output. Features selected for training network is given in Table 8
A NN based system takes the input in numerical form; hence the word features are converted into suitable numerical value. Mapping of features into numerical values is shown in Table 9.

Parameter	Feature type
Candidate suffix	Morphological feature
Candidate word category	Syntactic feature
Verb suffix	Morphological feature
Verb prefix	Morphological feature
Verb root	Lexical feature
Successive word	Context based feature
Previous word	Context based feature
Probability vector for suffix	Corpus based feature

Table 8: Feature selected for NN training

Category	
1	Pronoun
2	Noun
3	Verb
4	Avyaya
5	Adverb

Prefix(verb)	
1	up
2	anu
3	adhi
4	a
5	abhini

Verb root			
1	gachch	11	dhuh
2	pach	12	yAch
3	kar	13	danD
4	bhu	14	rudh
5	anay	15	prichch
6	As	16	Chi
7	pashy	17	bru
8	vas	18	Ji
9	dIy	19	math
10	yachch	20	suSh

Vibhakti-karka	
1	Agent
2	Object
3	Inst
4	Cause
5	From
6	Relation
7	Location

Suffix(verb)	
1	ti
2	si
3	at
4	Ami
5	anti
6	AmH
7	te
8	tH

Suffix(noun)	
0	No suffix
1	Am
2	H
3	Aya
4	ebhyH
5	Asya
6	en
7	e

Table 9: Sample set of encoded values

Input coding algorithm reads the pre annotated text and generates the data for training the neural network.

BPN is feed forward multi layer network consisting of mainly three layers. Algorithm uses two passes - forward and backward pass. In the forward pass, inputs are multiplied by respective weight and a bias added to it. Weighted sum of input along with bias is fed as input to hidden layer. Hidden layer uses a squashing function to limit the output value in desired range. Output from hidden layer is multiplied by respective weight and fed to output layer. Sigmoid function is used to limit the range of output. Output obtained is compared with actual value and error is calculated as difference of the two. Error is a measure of difference between actual and the desired output. This calculated error is propagated back in the backward pass. To improve the performance of the network, weights are modified as a function of propagated error. In the forward pass, weights of the directed links remain unchanged at each processing unit of the hidden layer. For n input values weighted sum is obtained and sigmod function is applied to this weighted sum.

Time taken to train the network is directly proportional to size of data. If number of neurons is increased, training time increases. Classification of *pAninI kAraka* with NN require large size corpus for training. Hybrid model overcomes the problem of large training time by classifying the word with their *vibhakti* value in non-conflicting situations and applying NN under conflicting situations only. This requires a small set of data and network is trained for conflicting classes only, thereby reducing the time. All the cases under same conflicting domain require same network. Major conflicting domains are (1, 2) (3, 4, 5) (6, 7).As data set for each conflicting case focuses on limited set of suffix, small data size is sufficient. For example *am* ending suffix fall in two class 1 and 2; NN was trained on these cases and result of the cases is presented next section.

7. RESULT AND DISCUSSION

Training time for each conflicting domain is reported in Table 10.

***pAninI* class**	**F**	**n**	**C**	**P**	**R**	**k**
Karta	122	110	108	0.981	0.885	0.930
Karma	64	61	58	0.951	0.906	0.931
Karan	52	50	49	0.980	0.942	0.957
sampradAn	35	30	29	0.966	0.828	0.891
apAdAn	34	30	32	0.93	0.823	0.873
Adhikaran	38	35	33	0.943	0.868	0.903

Table 10: Performance of NN for Various Data Set Training Size=50 Test Data Size=10

Fifty sentences used in training phase and ten sentences in testing phase. As depicted in the Table 6.1; 90% accuracy is achieved in *am* and *ebhyH* domain. Training of network for *am abhyam* conflicting case is given in Fig 8 and Fig 9.

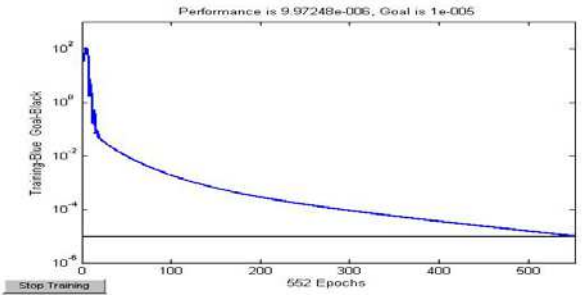


Figure 8: Training graph for *am* data set

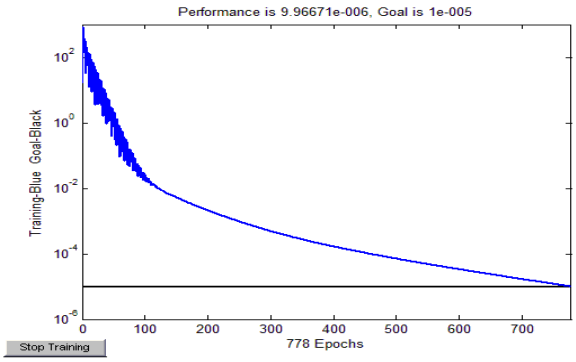


Figure 9: Training data for *abhyam* data set

After training the network for conflicting cases; algorithm is tested on 100 sentences and accuracy of the output obtained is calculated by finding the F-score as given in Eq 6.1

$$F\_score = 2(p \times r)/(p+r) \quad \text{---------- (6.1)}$$

It uses precision (p) and recall (r) to compute the score.

Precision (p) = Number of correct result / Number of returned result

Recall (r) = Number of correct result / Number of results that should have been returned.
F_score is understood as weighted average of precision and recall and is calculated as given in Eq (6.1). F-Score of each class under NN is given in Table 11.
f= frequency of occurrence
n= number present in the data
c=number correctly identified

Suffix	Conflicting	Training time	Epochs	Correctly identified
am	1, 2	09.89	125	9
Abhyam	3, 4, 5	11.27	210	8
ebhyH	4, 6	08.54	116	9
e	1, 2, 7	10.32	198	7

Table 11: Result for Hybrid Classification

Hybrid model is better approach for semantic classification as compared to pure rule based system or NN based system. Performance of NN is dependent on the size of annotated corpus available for training with good coverage of the vocabulary and suffix. As it is suffix driven analysis, annotated corpus must include the suffix attached to the words. Due to lack of availability of corpus, with good coverage, results of pure NN based system lags behind hybrid system. Hybrid model exploits the potential of rules of the grammar and handles conflicting situation by implementing NN model. Requirement of large size corpus is reduced as corpus is designed for conflicting cases only. Pure rule base system require in depth knowledge and understanding of complex set of recursive and meta rules for transformation and exceptional cases. A person with good computational skill along with complete *pAninian* knowledge is difficult to achieve.
Case frames for 100 sentences are generated by three algorithms and accuracy is checked at word level and sentence level. Word level accuracy is correctness of semantic tag assigned to each word and sentence level accuracy is correctness of generated case frame. Word level accuracy is discussed in Table 5.3, 5.6 and 5.8. For sentence level accuracy, accuracy of case frame is calculated. Accuracy of case frame depends upon two parameters:-

- Number of significant words from a sentence appearing in case frame ---(x)
- number of words tagged correctly ----(y)

x \ y	100%	50-100%	50%	Total
100%	56	11	5	72
50 -100%	12	5	2	19
< 50%	8	1	--	9

For hybrid model, sentence level accuracy is calculated and presented in Table 12.

Table 12.: Case Frame Accuracy for 100 Sentences under Hybrid Model

Out of 72 CF with all significant word of sentences; 56 are correctly tagged giving the accuracy of 77 % which is so far the best as none of the NLP processor for Sanskrit language has worked on KR tool generation for Sanskrit language.
Use of NN in NLP is less frequent due to complexity prevailing in both domains [1]. Sanskrit language has rich inflectional morphological structure suitable for computational processing. Tabular declension of words with syntactic-semantic significant suffix occupying predefined cell position drives the path for well structure knowledge representation mechanism. Identifying the semantic class of the word with suffix driven analysis under *pAninI* concept was the main theme of the work. Use of NN for resolving conflicting *kAraka* role under *pAninI* framework appears to be a better mechanism for semantic labeling of words. Initial identification is a baseline model upon which further extensions can be developed. Enhanced corpus with better coverage can further improve the results.

REFERENCES

[1]. Babu, A. Suresh, K. & Pavan, P.N.V.S. 2010. *Comparing Neural Network Approach With N-Gram Approach For Text Categorization.* International Journal on Computer Science and Engineering. 2(1): 80-83.

[2]. Bharati, A. & Kulkarni, A. 2007. *Sanskrit and Computational Linguistic*. First International Sanskrit Computational Symposium. Department of Sanskrit Studies, University of Hyderabad.

[3]. Bharati, A., Kulkarni, A. & Sivaja S. N. 2008. *Use of Amarako´sa and Hindi WordNet in Building a Network of Sanskrit Words.* 6th International Conference on Natural Language Processing.-ICON-08.Macmillan Publishers. India.

[4]. Brants, T. 2000. *TnT - A Statistical Part Of Speech Tagger*. Proceedings of the 6th Conference on Applied Natural Language Processing 2000.

[5]. Briggs, R. 1995. *Knowledge Representation in Sanskrit and Artificial Intelligence*. AI Magazine, 6 (1): 32-39.

[6]. Dwivedi, K. (padamshri) 2002. *Prarambhik RachanAnuvAdakumaudi*. VishavavidyAlaya PrakAshan. Varanasi. 19th ed. ISBN:81-7124-86-0

[7]. Hellwig, O. 2007. *SanskritTagger, a stochastic lexical and POS tagger for Sanskrit.* First International Sanskrit Computational Linguistics Symposium. LNCS Springer. 5402:266-277.

[8]. Huet, G. 2003. *Towards Computational Processing of Sanskrit.* Recent advances in Natural Language Processing. Proceedings in International conference ICON.:1-10

[9]. Hyman D. M. 2009 *From pAninian sandhi to finite state calculus.* Sanskrit Computational Linguistics Springer-Verlag.

[10]. Jha, G. & Chandrashekar, R. 2010. Annotation Guidelines for tagging Sanskrit using MSRI-JNU Sanskrit tagset http://sanskrit.jnu.ac.in/corpora/ (Browsing Date: 25th Dec 2011).

[11]. Kak, S. C. 1987. *The panian approach to natural language processing.* International Journal of Approximate Reasoning. Elsvier Publishing. 1(1):117-130.

[12]. Kiparsky, P. 2002. *On the architecture of P¯anini's grammar*. International Conference on the Architecture of Grammar.

[13]. Kumar, D. & Josan, G. 2010. *Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey*. International Journal of Computer Applications. 6(5): 32-41.

[14]. Mittal, V. 2010. *Automatic Sanskrit Segmentizer Using Finite State Transducers.* LRTC, Research Centre. IIIT-Hyderabad. Proceedings of the ACL 2010:85-90.

[15]. Pedersen, M., Eades, D., Amin, S. K. & Prakash, L. 2004. *Relative Clauses in Hindi and Arabic: A Paninian Dependency Grammar Analysis* in COLING.

[16]. Rao, B. N. 2005. *Panini and Compuer Science – into the future with knowledge from past.* A Sourcebook:9- 13.

[17]. Selot, S. & Singh J. 2007. *Knowledge representation and Information Retrieval in PAninI Grammar Framework.* International Conference ICSCIS- 2007. 2: 45-51.

[18]. Selot, S., Tripathi, N. & Zadgaonkar, A. S. 2009. *Transition network for processing of Sanskrit text for identification of case endings*" icfai Journal of Computer Science 3(4):32-38.

[19]. Shan, H. & Gildea, D. 2004. *Semantic labelling by Maximum entropy model.* University of Rochester Technical Reort 847.

[20]. Timothy, J. D., Hauser M. & Tecumseh, W. 2005. *Using mathematical models of language experimentally* TRENDS in Cognitive Sciences 9(6):284-289.

[21]. Vaidya, A., Husain, S., Mannem, P. & Misra, D. S. 2009. *A Karaka Based Annotation Scheme for English* Language Technologies Research Centre. IIIT Hyderabad. Springer-Verlag: LNCS 5449:41–52.

E-Licensing in DGFT: A Best E-Governance Application

V. S. Rana

Submitted in October 2012, Accepted in March 2013

Abstract - This is a case of effective and efficient e-Governance where the licenses are issued electronically by using of IT for web based delivery service in the Directorate General of Foreign Trade (DGFT). Now a days complete licensing procedure is dealing through electronically i.e. Exporter/Importer apply electronically to get the Importer Exporter Code(IEC) , submitting the fees as well as necessary documents electronically and received the IEC electronically. Trading Community is also availing the Online facility to submit the application for any licensing scheme, depositing the licensing fees, enclosing the required document from their end. The official procedure is also automated like initiating the note sheet, generating the ecom number, consolation of license fees and issuing the license to the exporter. There is fifty percent role of customs involved in trading. So that Electronic Data Interchange(EDI) facility is also established with Customs. In addition to above services, Bank Realization Certificate(BRC) is also integrated with this system. Henceforth an Exporter/Importer is equipped with electronic services without visiting to the office of DGFT, Customs and Bank. The web has been strategically leveraged for reengineering and transformation of trade processes for an economic trade facilitation.

Index Terms - E-Licensing, DGFT, E-Governance

1. INTRODUCTION

NIC-DGFT (Commerce and Industry Informatics Division) playing a significant role in architecting & implementing e-Governance initiatives with the best possible technology support in the Directorate General of Foreign Trade(DGFT). DGFT is a country wide organization and responsible to increase the export of the country has been discussed in [12]. Appropriate backbone ICT infrastructure has been established in DGFT which includes OFC-based Internet connectivity with Gigabit-based Local Area Network (LAN), Video Conferencing, IT equipped help desk, etc. supported by a team of highly qualified IT professional. [1-11] are the various e-Governance models defining the various e-Governance indicators and parameters which have been implemented in different forms. 'Trade Facilitation' is a key determinant of a country's competitiveness in the international market so there was a thrust of traders to familiar with it . Over the years, Government of India has taken various initiatives to simplify and rationalize procedural complexities in exports in order to put in place an efficient and effective trade facilitation mechanism and reduce the implicit transaction costs associated with the enforcement of legislation, regulation and administration of trade policies involving several agencies such as Customs, Airport and Port Authorities, bank, trade ministry etc. The transaction cost has been evaluated at about 8 to 10% of the value of exports and any mitigation in this has a permanent benefit accruing to the exporters.

NIC-DGFT has played catalytic & significant role in implementing e-Governance project in the DGFT with an aim to leverage IT for transparency and better governance. ,Keeping in view the object Directorate General of Foreign Trade had set up an online trade facilitation system. It containing EDI interfaces with the Trade Partners and all concerned in the value chain have been established. Customs, Banks, Trade and Industry and other Government Agencies are the part of this mechanism. Electronic Data Interchange (EDI) is core driver for facilitating international trade and one of the key initiatives is electronic transmission of foreign exchange realization details on exports by banks on a daily basis under the Electronic Bank realization Certificate (e-BRC) initiative. Exporter will not be required to make any request to Bank for issuance of Bank export and Realization Certificate (BRC). This will establish a seamless EDI connectivity amongst DGFT, Banks and Exporters. This is significant step to reduce transaction cost to the exporters..

2. OBJECTIVES OF DGFT

The major objectives of DGFT are as follows:

- 2.1. Effective and efficient e-governance services,
- 2.2. Globally accessibility of the e-services.
- 2.3. Maintaining the integrity of public services.
- 2.4. Reduction in transaction cost and time.
- 2.5. Elimination of fraud practice of trade and industry
- 2.6. Physical visit of exporter of the office reduced to minimum.
- 2.7. Publishing of Monthly license Bulletin.
- 2.8. Implementation of single common document for the trade.
- 2.9. To move the DGFT in paperless environment.

3. ACIEVMENTS

To achieve the above said goals DGF organization requires intensive use of ICT infrastructure. The online service has become a core implementation strategy for delivery of an efficient, transparent and easy to access service. For the implementation of powerful and successful e-Governance complete setup has been renovated in the following manner as:

- 3.1. DGFT has been automated in all respect
- 3.2. DGFT web site prepared and hosting annually including latest policies, procedures, Circulars, Notifications and public Notice etc

National Informatics Centre, MCIT, Govt. of India, New Delhi, India; E-mail: vsrana@nic.in

3.3. Launching the web based application to get the licenses for the trading community
3.4. Creation of Central Data ware house of license data.
3.5. Global trade facility is available round the clock through the DGFT Portal http://dgft.gov.in
3.6. EDI with Customs is operational.
3.7. Net banking facility is made available to pay the licensing fees. MOU between DGFT and 43 banks has been signed.
3.8. Serving the trading community of 5.5. lakh exporters and importers using online facility 365x24x7.
3.9. Professionally managed help desk id operational at DGFT HQ as well at regional office.
3.10. All 36 DGFT port offices are providing the trading facilities country wide.

4. MAJOR PARTNERS OF TRADE

4.1 **Customs;** Message Exchange pertaining to various FTP schemes like Advance Authorizations (AA), Duty Exemption Passbook (DEPB), Export Promotion Capital Goods (EPCG) etc.
4.2 **Banks;** Message Exchange to obtain Foreign Exchange realization against exports (under implementation)
4.3 **Export Promotion Councils (EPCs);** Message Exchange / uploading of membership details of registered exporters (e-RCMC)

5. KEY TECHNICAL ATTRIBUTES OF DGFT'S ONLINE SERVICES

All processes and procedures have, therefore, been reengineered leveraging the web technology. Capability, flexibility and management of DGFT's website is vital to the process of trade facilitation.

The four major key attributes of the DGFT's website are:

5.1 A broad application Filing Spectrum
5.2 Security Features
5.3 Web Management
5.4 Technology

The above key attributes of the DGFT's website are indicated in the following schematic (Figure 1).

6. CITIZEN CENTRIC APPROACH

6.1 Web based operational environment is made available for Trade policy and procedure implementation globally, on 24x7x365 basis for all citizen.
6.2 DGFT Head Quarter with all 36 regional offices spread (but virtually being one) across the country for providing the online trading facility at user end.
6.3 e-Licensing facility for almost schemes like Advance Authorization (AA), Duty Entitlement Passbook (DEPB), Export Promotion Capital Goods (EPCG), Focus Product Scheme(FPS), Focus Market Scheme(FMS), Vishesh Krishi and Gram Udyog Yojna (VKGUY) Scheme, Status Holder Incentive Scrip(SHIS) Scheme, Market Linked Focused Product (MLFPS) Scheme, Served from India Scheme (SFIS) and Agri Infrastructure Incentive Scrip (AIIS) Scheme etc.
6.4 On the DGFT web site http://dgft.gov.in a facility has been provided to search/enquire about the current Import Policy of an item by entering either ITC (HS) Code of that item or brief description of that items. This would be of major help to trade and industry as well as to academicians and researchers.
6.5 Organization has undertaken a through revision of Foreign Trade Policy/ Handbook of procedures electronically to make it more user friendly. Substantial efforts have been made to remove ambiguities in language, delete repetitions and harmonize the text with amendments to policy and new policy announcements.
6.6 An extremely challenging and significant EDI initiative e-BRC has been launched by DGFT It would herald electronic transmission of Foreign Exchange Realization from the respective Banks to the DGFT,s server on a daily basis. In addition to this EDI linkages with Trade and Industry, Government. Agencies and related EDI community partners i.e., Customs, and EPC's etc. e-BRC would facilitate early settlement and release of FTP incentives/entitlements for the exporters/importers.

7. TRANSITIONAL COMPATIBILITIES

7.1 The 'on-line' filing facility is user friendly, data input through structured screens, access controlled by DSC's, inbuilt facility to edit and validate before submitting data and availability of FAQ's to assist filing.
7.2 Status of Authorization and Importer Exporter Code (IEC)
7.3 Electronic Fee Transfer (EFT)
7.4 Secure and automated EDI based environment with 'on-line' EDI Message Exchange with community partners
7.5 Covers all models of e-governance i.e. B2G, G2G, G2B, G2C and C2G.

8. SEARCH ENHANCEMENT

A comprehensive user friendly search facility is available on the web portal for the people to search any trade related information. Any Exporter/Importer may know the status of any Authorization as well as IEC at any time from any where. All Trade related documents may be obtained through the menu. Latest updates of Foreign Trade Policy and Procedure, RTI Related Information that who is who ? Citizen charter etc may be noted down from the DGFT site. All type of format may also be downloaded as and when required.

9. WEB SECURITY FEATURES

9.1 The user authorization is through digital signatures. However option to login through a user name and password also exists to provide flexibility

9.2 The Digital Signature includes embedded IEC details also which when registered on DGFT's website get validated and ensure high level of security. DGFT has also recently migrated to a 2048 bit encryption for higher level of security. DGFT is geared to handle any changes which may be required after implementation of interoperability in issuance of Digital Signature Certificates (DSCs)

9.3 Database and application server maintained under firewall

9.4 A three tier architecture used for the application

9.5 Physical security is ensured by NIC Data center authorities

10. MESSAGE EXCHANGE BEHAVIOR WITH VARIOUS COMMUNITY PARTNERS)

The message exchange behavior with various community partners may be shown as given in the table 1:

Network Partner	Projects/ Activities	Network Topology	Mode	Security	Message Exchange file format
Customs	Authorization, Shipping bill	One to one	Offline	Access control through DSC	Flat file through FTP
Banks(e-BRC)	E-BRC	One to many	Offline	Access control through DSC	XML file upload
EPC's(e-RCMC)	E-RCMC	One to many	Offline	Access control through DSC	XML file upload
Banks	EFT	One to many	Online	Access control through DSC	Integration with bank website

11. TECHNOLOGY ADOPTED

For smooth functioning of e-Governance project we have adopted the object oriented language (java) and supporting the database (DB2) at back end. The details of the technology tools are as follows:

11.1. J2EE technology (Applet, Servlet, Enterprise Java Beans, JSP, ASP),XML IBM DB2 as database with digital signature.

11.2. J Builder 2007, J2SDK/J2SEE tools for applications development.
IBM Web Sphere, Macro Media JRun Web Server for application servers.

11.3. Rational Suite is implemented for documenting/ designing/development of the application.

11.4. The website is being updated using the Extended Markup Language (XML) technology.

11.5. Whole DGFT Organization is connected with internet / intranet / VPN through very high speed connectivity with NICNET infrastructure.

The continuous technology up gradation has not only prevented obsolescence but has kept our infrastructure robust from security, capability, flexibility and compatibility perspective

12. EDI/ONLINE FILLING ERROR RESOLUTION SYSTEM

12.1. EDI Help Desk is manned by expert professionals

12.2. managed by the EDI Division to resolve EDI related complaints. Nodal officers have been nominated at DGFT / major RA's to monitor / resolve EDI related complaints from trade and industry.

12.3. A tracking system has been established on the basis of a unique complaint number. An online complaint registration system is implemented.

13. OTHER IMPORTANT INFORMATION LINKS

13.1. Right to Information Act (RTI)

13.2. Citizen Charter

13.3. DGFT's Regional Offices, Ministry of Commerce & Industry, Directorate General of Commercial Intelligence & Statistic (DGCI&S), Central Board of Excise & Customs (CBEC), Special Economic Zone (SEZ), World Trade Organization (WTO), Customs Port Location Code

13.4. Public Grievances

14. ECONOMIC OUTCOMES

Due to providing the SMART services by e-Governance to Trading Community the money is saving by each stake holders. Public as well as Government is benefited with this application. The major factors as below:

14.1. Application for licensing can be filed from any where

14.2. Status of the application can be tracked just click a button

14.3. Physical visit of exporters of the office has been reduced to minimum

14.4. Interview through Video Conference saves on an average Rs. 50,000/per interview
14.5. Cost of stationery brought down up to 80%
14.6. Reduction in paper work due to paperless operation in DGFT
14.7. Reduction in transaction cost
14.8. File preparation cost come down to 0

15. TIME FACTOR OUTCOME

15.1. Licensing application preparation time come down from 5 hours to 5 minutes
15.2. Application processing time has also come down from 45 days to 5 hours.
15.3. Message exchange time for license verification has come down from 6 month to instant
15.4. Status of application can be traces on urgent basis.
15.5. Complete process is very fast
15.6. Collection of license fees as well as consolation is too fast due to EFT implementation .
15.7. Reduction in transaction time

16. GENERAL OUTCOME

16.1. Fraud practice of trade and industry in eliminated
16.2. Entire process is transparent
16.3. Collection of license fees is easy and systematic
16.4. G2G, G2B, G2C, C2G, B2G model of e-Governance
16.5. Secured automated EDI based environment,
16.6. Implementation of single common documents

17. E- LICENSING IN VIEW OF RESEARCH TECHNOLOGY

In India, over the last two decades, Information and Communication Technology(ICT) has emerged an effective tool to deliver services to the people. Expansion of Telecommunication Infrastructure and penetration on Internet in large parts of county, has enabled the government to provide effective, efficient and multichannel delivery of government services to the citizens. Initially the emphasis of e-governance initiated towards G2 G services relating to automation and computerization of inter functioning of the government since last few years focus on e-governance has shifted to electronic delivery of services to the citizens at his end. So that as the interest in new and expanded e-governance increases public managers find themselves making decisions about information and information technology for which they are often unprepared or ill-equipped. Identification of the complexity and risk of IT decisions public managers involved in making these types of decisions has spurred the development of many structured tools and rigorous to support IT business case analysis and risk assessment strategies recommended in some government agencies and required in other also as referred in [11].

A gap analysis between a selected set of practitioner tools and a set of key success factors of IT initiatives has the potential to inform questions about the relationship between research and practical. A gap analysis strategy represents an opportunity to do a component-by-component analysis to determine the extent to which the decision of each reflects awareness of relevant research on information system success. The gap analysis is comprised of the steps as follows:

17.1. A review of current literature in information system research is used to identify factors found to influence the success of IT initiatives.
17.2. The research identified and described a set of tools used for government IT initiatives. These tools to be selected based on their visibility and central role in informing practitioners at the National Level.
17.3. A comparison of the factors against the selective descriptions was conducted
17.4. An identification of the gap between the research and the practical tools is presented and discussed.

18. EMERGING CHALLENGES FOR E-LICENSING

Although providing numerous opportunities for better governance globalization and ICT have also brought in many new challenges for **e-Licensing** like information and data, information technology, organizational and managerial, legal and regulatory, institutional and environmental factor etc. The major challenges may be classified in a following manner:

18.1. **Information and data challenges:** e-Licensing initiatives are about the capture, management, use, dissemination, and sharing of information. A number of challenges relate to the information that is at the core of e-Licensing initiatives. Information and data quality, security issues, Technological incompatibility, Technology complexity, Technical skills and experience, technology newness, project size, management attributes and behavior, organizational diversity, lack of alignment of organizational goals and project multiple or conflicting goals, restrictive laws and regulations, intergovernmental relationship, budget and political pressure, autonomy of agencies etc. are the major challenges to implement the e-licensing application.
18.2. **Information Technology:** Technology incompatibility has also been identified as one difficult challenge to **e-Licensing** project. Very different and old systems increase complexity of IT projects, complexity and newness of technology are also constraints to effect the result of IT projects. The lack of relevant technical skills as well as the shortage of qualified technical personnel within the project team has been found to be an important challenging factor.
18.3. **Organizational and managerial:** The size of the project and the diversity of the users and organizations involved are two of the main challenges of **e-Licensing** project. There are lack of alignment between organizational goals and the existing project, secondly individual interests and associated behaviors lead to resistance to change internal conflicts.
18.4. **Legal and regulatory:** Like most of government department DGFT is also created and operate by virtue

of a specific formal rule or group of rules. In making any kind of decision, including those in this project, public managers take into account a large number of restrictive laws and regulations.

19. RECOMMENDED RESEARCH METHODOLOGY TO OVERCOME THE CHALLENGES

To achieve success in e-Licensing as e-governance initiative a set of strategies may be drawn by mapping the challenged categories. This illustrates the degree of correspondence in research itself between challenges and possible strategies for meeting those challenges as:

19.1. **Information and data strategies:** Information and data challenges require a overall plan for managing data and information produces. A quality and compliance assurance program is an effective strategy for dealing with information and data challenges managers have attempted to minimize data related problems by sharing standards, definitions and meta data with their potential partners like customs, banks, export promotion councils etc. In spite of this continual feedback from partners and users should maintain.

19.2. **Information Technology Strategies:** IT related issues i.e. ease of use, usefulness, demonstrations and prototypes etc. are success strategy. Well skilled and respected IT leader, expert project team, clear and realistic goal, identification of relevant stakeholders and user involvement proper planning, good communication, clear milestones and measurable deliverables adequate funding, best practice review, IT policies and standards etc. are the key success strategies.

19.3. **Organizational and Managerial Strategies:** For the successful IT initiatives there is a clear realistic goals is an important factor. Relevant stake holders and getting their involvement in the project development process, specially end users has also been found to be an effective strategy in overcoming the organizational and managerial challenges. Strategic planning technique can be seen as an umbrella for more specific strategies such as milestone and measurable deliverables, good communication channels. It is also extremely important to take care of developers and end users current skills and training needs. Successful projects need a balanced combination of technical managerial skills and expertise among their members.

19.4. **Legal and regulatory Strategies:** Restrictive laws and regulations developed prior to or in ignorance of technologies relevant to **e-Licensing** can affect the success of project. Our strategy for responding to these challenges is to invest in changes to the regulatory environment that allow for or enable adoption of emerging technologies. As Digital Signature Technologies for example required statutory changes in most jurisdictions before they would be adopted for use. Developing appropriate government wide IT Policies and standards can also provide and adequate framework for e-government initiatives to be successful.

20. FUTURISTIC RESEARCH TOOL

e- Licensing is a key challenge for government today as they involve multiple stake holders and multiple processes and demand considerable co-ordination and collaboration as well as managerial and financial resources we may adopt the following strategies as:

20.1. Promoting advance ICT training, education and research as and when conception of new technologies.

20.2. Negotiating and influencing the proper adoption of international frameworks, norms and standards by participating actively in the governance of the global information economy.

20.3. Documenting best success and worst failure benefiting knowledge

20.4. Promoting innovation and risk taking through fiscal concessions and availability of venture capital, creating an investment climate for domestic and foreign investment in ICT sector

20.5. Developing a supportive framework for early adoption of ICT and creating a regulatory framework for ICT-related activities, e.g. fixed and mobile communication, e-commerce and internet services.

20.6. Application of Online Performance Tracing System

20.7. Implementation of online Audit System.

20.8. Integration of Realty simple syndicatio (RSS) system with existing system for wider level simplification.

20.9. Inclusion of Cloud computing concept as futuristic approach.

20.10. Adoption of Yi Fi communication in the entire organization.

20.11. User requirement analysis is a major tool for refinement of the project

20.12. Use feedback analysis is also a powerful key factor for improvement of project.

20.13. Cost analysis is always a considerable measure for the project.

21. CONCLUSION

In this paper, I have presented the effects of e-Governance indicators in Directorate General of Foreign Trade, Ministry of Commerce and Industries, Govt. of India that the trading community availing the maximum facilities in minimum time from their end only within transparent environment. The Government of India, department of Electronics and Information Technology, has initiated national e-governance plan for the execution of e-governance projects in the country. In the same pattern we have applied the e-Governance module in DGFT to move in a paperless Journey. The various outcomes are indicated to support the effective and successful e-Governance.

This is the case study of best e-governance project. This project is highlighted in various e-governance seminar /workshop. This is the first govt. project in which ICT was implemented with digital signature and electronic fund transfer facility in 1998. Now a days this office is operational in paperless environment.

REFERENCES

[1]. E-Governance Initiative of the Government of Maharashtra available at: http://www.maharashtra.gov.in

[2]. Impact Assessment Study of E-Government Projects in India available at: http://www.iimahdqw2007.ernet.in/egov/documents/impact-ass

[3]. E-Governance Initiative of the Government of West Bengal available at: http://www.westBengal.gov.in/it_policy_egovernance.htm.

[4]. E-Governance in state of Madhya Pradesh available at: http://www.mpgovt.nic.in

[5]. E-Governance Centre at Haryana Secretariat available at: http://www.expressindia.com

[6]. E-Governance in Andhra Pradesh available at: http://www.ap_it.com/egoverenance.htm

[7]. Mechanism of single window clearance system available at: http://www.rajgov.org/news/single window.htm

[8]. E-Governance in Himachal Pradesh available at: http://www.economictimes.com.

[9]. E-Governance in Ministries/Departments and State Governments available at http://www.mit.gov.in/eg.ms.asp

[10]. Sanjay Dhingra, "Measuring IT effectiveness in Banks of India for sustainable Development." BIJIT - BVICAM's International Journal of Information Technology, July-December 2011, Vol.3 No.2

[11]. J. Ramon Gil-Garcia, Theresa A. Pardo. " E-government success factors: Mapping practical tools to theoretical foundations" Government Information Quarterly 22(2005) 187-216 USA

[12]. Ajay Kumar Gupta, Kishore Kumar and Madaswamy Moni," Micro Irrigation Census Computerization: A step towards ICT for micro level planning in Water Resources Management and Planning to achieve Rural Prosperity", BIJIT - BVICAM's International Journal of Information Technology, July-December, 2010, Vol.2, No.2

[13]. V. S. Rana, "An Innovative Use of Information & Communication Technology (ICT) in Trade Facilitation in India", BIJIT - BVICAM's International Journal of Information Technology, July-December, 2012, Vol.4, No.2

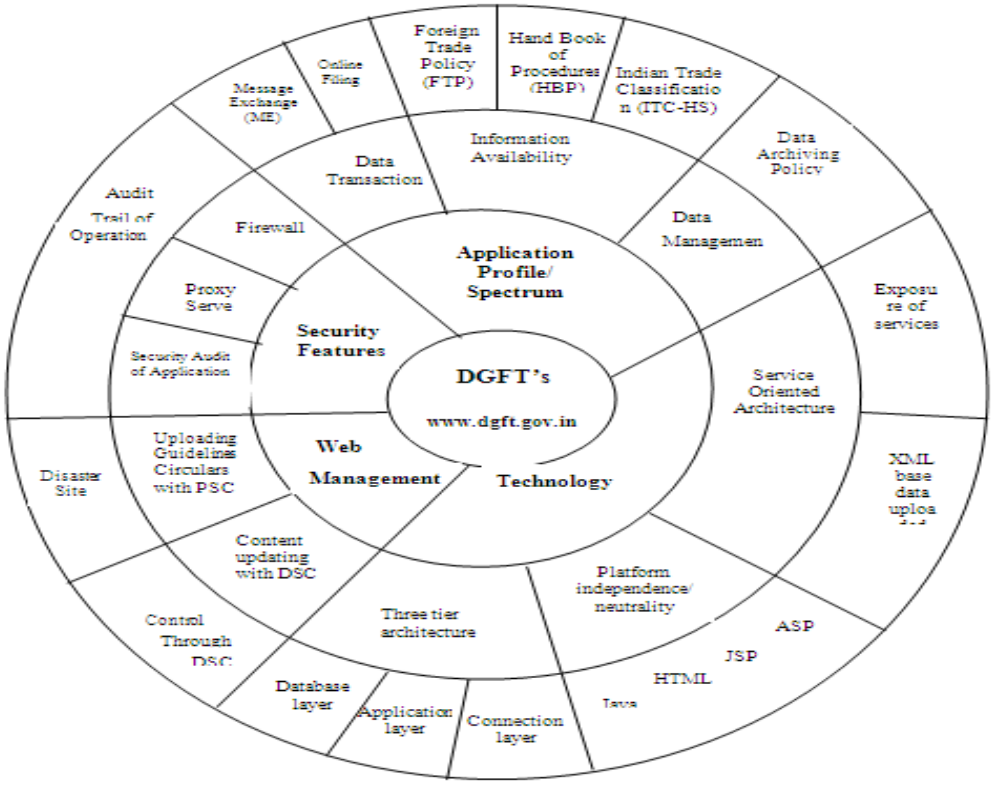


Figure 1: Key Attributes of DGFT Website

Miniaturisation of WLAN Feeler Using Media with a Negative Refractive Index

Bimal Garg[1] and **Ranjeet Pratap Singh Bhadoriya**[2]

Submitted in October 2012, Accepted in March 2013

Abstract - It presents a rectangular microstrip patch antenna integrated with combination of pentagonal and hexagonal shaped structure etched at the height of 3.276 mm from the ground plane. It is demonstrated that the application of the media with a negative refractive index or metamaterial eliminates the spurious harmonics (these are those unwanted dips which shows in the S11 graph) associated with the original structure. Furthermore the return loss is improved by the inclusion of the metamaterial structure reaching -27.1919 dB compared with -10.1286 dB achieved by the original patch antenna structure alone. Main focus in this design process is not to reduce the return loss but reduce the size of the antenna and this target has been achieved by reducing the size of antenna up to 65%. Numerical simulation results show that this proposed design possesses several desirable characteristics, for instance, high bandwidth, low loss and improved directivity compared to the alone RMPA. The CST-MWS software is used for designing and simulation, and MS-Excel for metamaterial proving.

Index Terms - Media with negative refractive index (metamaterial), rectangular microstrip patch antenna (RMPA), permittivity, permeability, NRW approach, Return Loss.

1. INTRODUCTION

In last decade the peremptory of Wireless communications systems have grown drastically. To fulfil this requirement, multifunction antennas have been designed for multipurpose operation over different wireless services. Recent improvement in communication technology and extensive growth in the wireless communication market and user demands exhibits the need for compact, reliable and efficient, wireless systems. Integrating whole transmitter and receiver system on a single chip [1], [3] is the imagination for future wireless systems. This particular idea has the benefit of cost reduction and enhancing system reliability. Antennas have always been considered as the largest components of integrated wireless systems, consequently antenna miniaturization became a necessary piece of work in achieving a favourable design for integrated wireless systems. Moreover, compactness is important aspect in wireless communication, addition with the other parameters improvement like directivity, return loss, bandwidth [2]. These characteristics can be achieved by covering of microstrip patch antennas with metamaterial structures [4], [5].

[1,2] *Dept. of Electronics Engineering, Madhav Institute of Technology & Science, Gwalior, India*
[1]*r.pratap7872@gmail.com and* [2]*bimalgarg@yahoo.com*

Several researchers have been trying from years to reduce the size of the antenna. It has been attempted in many ways and different concepts were proposed. Recently, metamaterial based structure, originally proposed by Pendry, has opened the door to new design strategies, where miniaturization and compatibility in planar circuit technology are key aspects. In 21st century split rings resonators (SRRs), originally proposed by Pendry [6], [7], have attracted a great interest for the design of negative permeability, negative permittivity and left-handed (LH) effective media [5].

In late sixties (1967) Victor Georgievich Veselago [5], a Russian physicist was the first researcher who presented the theory of metamaterial, which exhibit negative permittivity ε, and permeability μ [16] also known as media with a negative refractive index or left handed material [11], [13]. In such a material, he showed that the phase velocity would be anti-parallel to the direction of Poynting vector. This is contrary to wave propagation in natural occurring materials. In the beginning of 21st century, papers were published about the first demonstrations of an artificial material that produced a negative index of refraction (that was discussed in last paragraph). By 2007, research experiments which involved negative refractive index or metamaterial properties had been conducted by many groups.

2. DESIGN METHODOLOGY

All the design work and simulation work has been done on the computer simulation technology microwave studio (CST-MWS). And the proving of the metamaterial which used to enhance the property of RMPA, Microsoft excel software is used. Initially dimensions were calculated for the operating resonant frequency i.e. 2.05GHz by using formulas shown below.For calculation of width and length of the patch antenna:

$$W = \frac{1}{2f_r\sqrt{\mu_0\varepsilon_0}}\sqrt{\frac{2}{\varepsilon_r+1}} = \frac{c}{2f_r}\sqrt{\frac{2}{\varepsilon_r+1}} \quad (1)$$

$$L = Leff - 2\Delta L \quad (2)$$

Where,

$$Leff = \frac{c}{2f_r\sqrt{\varepsilon_{eff}}} \quad (3)$$

$$\frac{\Delta L}{h} = 0.412\frac{(\varepsilon_{eff}+0.3)\left(\frac{W}{h}+0.264\right)}{(\varepsilon_{eff}-0.258)\left(\frac{W}{h}+0.8\right)} \quad (4)$$

$$\varepsilon_{eff} = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2}\left(\frac{1}{\sqrt{1+\frac{12h}{w}}}\right) \quad (5)$$

In above used formulas the symbols have their usual meanings.

e.g.

c = Velocity of light in free space,

ε_r = Substrate's Dielectric constant,

ε_{eff} = Effective dielectric constant,

L_{eff} = Effective length.

After dimension calculation design work has been done. Perfect electric conductor was used to make the patch antenna over the ground which also having the same material with substrate between patch and ground. RMPA at 2.05GHz frequency is shown in figure 1.

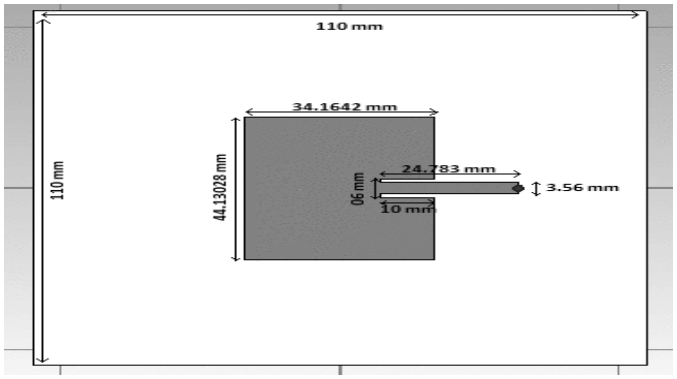


Figure 1: RMPA at height of 1.6mm from ground of 2.05GHz.

The simulation result of the patch shown in figure 1 is in graphical form shown in figure 2, with the return loss and bandwidth of -10.1286dB and 7.7MHz respectively.

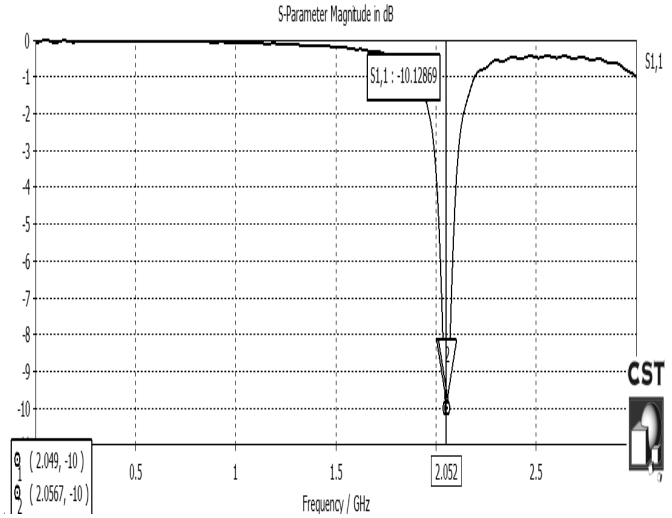


Figure 2: Simulation result of the RMPA with return loss of -10.12dB and bandwidth of 7.7MHz at 2.05GHz.

After the RMPA simulation the metamaterial cover is implemented over the patch antenna at the height of 3.2mm from the ground. The proposed metamaterial structure implemented as the cover of antenna with its dimension used in the proposed design is shown in figure 3.

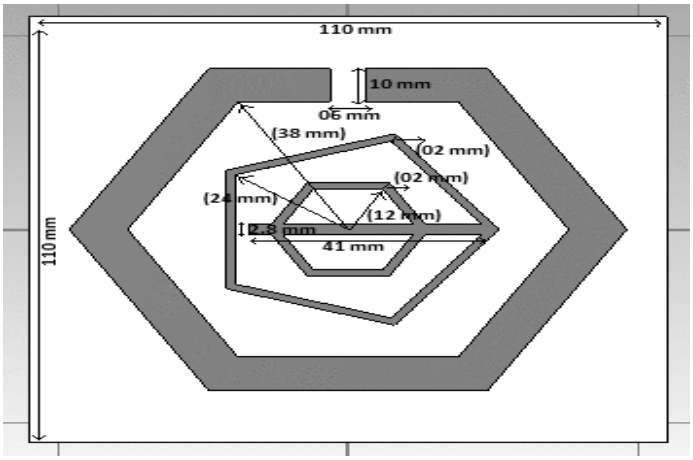


Figure 3: Proposed metamaterial structure at the height of 3.2mm from ground.

The simulation result after the implementation of the metamaterial over the rectangular microstrip patch antenna at the height of 3.2mm from the ground enhance the property of the RMPA alone and reduces the size of the antenna by shifting the lowest dip to a frequency other than the operative frequency i.e. at 0.651GHz. The size is being reduced to 65%. The simulation result with the metamaterial is shown in figure 4.

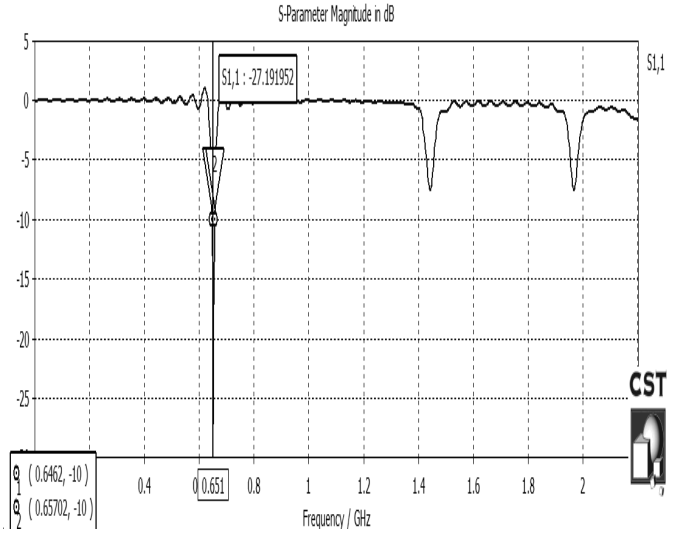


Figure 4: This simulated result is showing the return loss of -27.19dB and bandwidth of 10.82MHz at 0.651GHz.

Comparison of dimensions between reduced patch antenna using media with negative refractive index at operating frequency 2.05GHz and RMPA alone at 0.651GHz is in tabular form below.

	Dimensions of RMPA alone at 0.651GHz	Dimensions of RMPA using metamaterial works at 0.651GHz	Unit
Length	110.9574	34.1642	mm
Width	141.5426	44.1302	mm
Cut width	20	6	mm
Cut length	35	10	mm
length of feed	85.2926	24.7830	mm
Width of feed	14	3.56	mm

Table 1: Comparison of Dimensions

After comparing it is necessary to prove that the material here used to reduce the size of RMPA is Meta, NRW (Nicolson Ross Weir) approach [14] is used to prove it. The following formulas belong to NRW approach:

$$\mu_r = \frac{2.c(1-v2)}{\omega.d.i(1+v2)} \quad (6)$$

$$\varepsilon_r = \mu_r + \frac{2.511.c.i}{\omega.d} \quad (7)$$

Where,

V_2 = S21 - S11 or Voltage Minima,

ω = Frequency in Radian,

d = Thickness of the Substrate,

c = Speed of Light,

μ_r = Relative permeability,

ε_r = Relative permittivity.

In NRW approach, proposed design of patch antenna having metamaterial structure placed between two waveguide ports on both sides of antenna on X-axis to calculate S11 and S21 parameters. Y and Z planes are defined as the perfect electric and magnetic boundary respectively. Following that, the wave was excited toward the port 2 from port 1 or left to right.

Later on, after the simulation in CST-MWS software the S11 and S21 parameters were exported to MS Excel software for further calculation. In MS Excel equation number (6) & (7) were used for proving of structure that it is metamaterial. The result obtained using NRW approach are showing negative permeability and permittivity in figure 6 & 7 respectively.

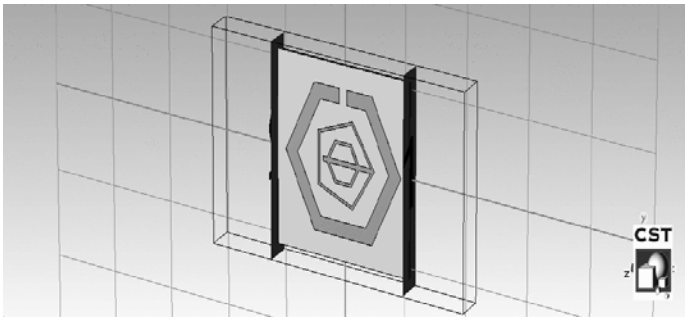


Figure 5: Proposed metamaterial structure between waveguide ports.

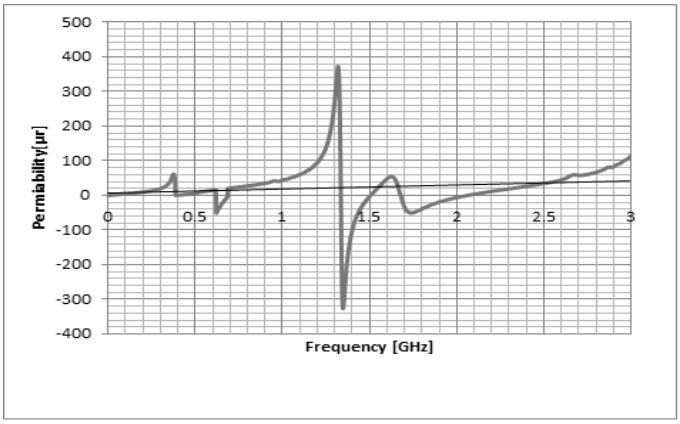


Figure 6: Permeability versus frequency graph obtained from Excel software.

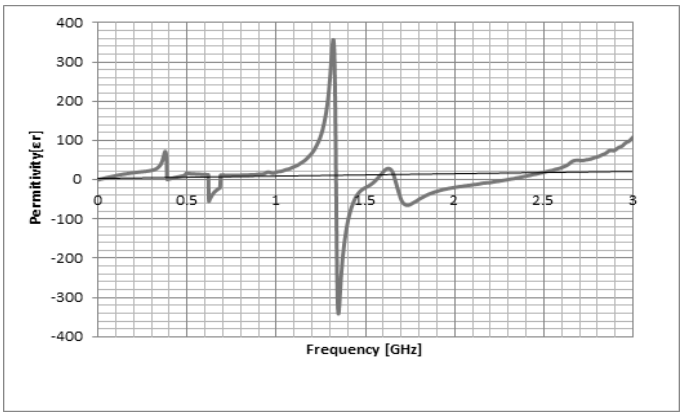


Figure. 7: permittivity versus frequency graph obtained from Microsoft Excel software.

The Table's generated for permittivity and permeability by using MS-Excel Software was too large, therefore the Table 2 & Table 3 shows the negative value of permittivity and permeability only in the frequency range 0.6419-0.6539GHz.

Frequency [GHz]	Permeability[μr]	Re[μr]
0.6419999	-31.9316370277838-14.5648307462409i	-31.93163703
0.64499998	-29.5759201229285-14.3467942395896i	-29.57592012
0.648	-27.3064654388456-14.1729215593206i	-27.30646544
0.6509999	-25.1190003040519-14.0363460044533i	-25.1190003
0.65399998	-23.0080937796368-13.9305290718914i	-23.00809378

Table 2: Sampled Values of Permeability at 0.651GHz Calculated on MS Excel Software.

Frequency [GHz]	Permittivity[εr]	Re[εr]
0.6419999	-37.1552405011718-24.6492429073745i	-37.1552405
0.64499998	-35.314195613912-25.2329994815107i	-35.31419561
0.648	-33.6325777383075-25.8021819587365i	-33.63257774
0.6509999	-32.0899516273428-26.3433936840906i	-32.08995163
0.65399998	-30.665474813861-26.8496952334226i	-30.66547481

Table 3: Sampled Values of Permittivity at 0.651GHz Calculated on MS Excel Software.

After proving of metamaterial it has been defined that the proposed structure to miniaturize the antenna was metamaterial. Post proving, hardware of the proposed design was constructed and analyzed using spectrum analyzer and the results of RMPA alone and incorporated feeler were compared. Figures are shown below.

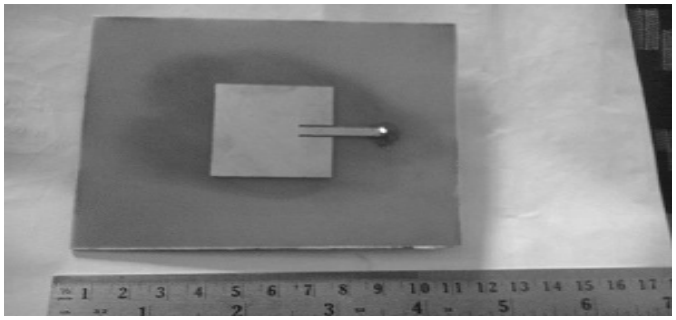


Figure 8: Hardware of RMPA alone at 2.05 GHz.

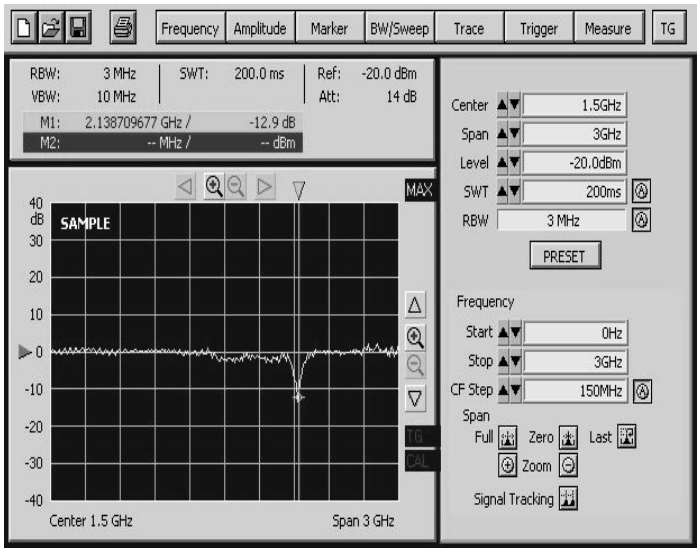


Figure 9: Analyzed result of patch showing return loss of -12.9 dB at 2.13 GHz.

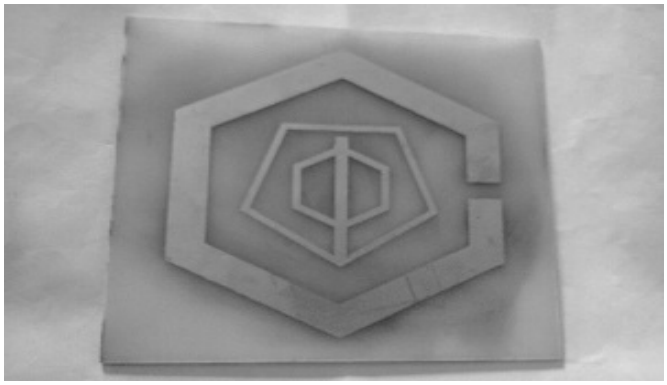


Figure 10: Incorporated metamaterial structure over patch surface.

3. CONCLUSION AND FUTURE SCOPE

By emphasizing RMPA with the Metamaterial structure the frequency on which it shows its maximum power output or lowest return loss is 0.651GHz. Table 1 shows the comparison of patch antenna designed at the frequency of 0.651GHz and at 2.05GHz with metamaterial. RMPA at 0.651GHz consumes a large area instead of RMPA at 2.05GHz. By using metamaterial

it became possible that the antenna at 2.05GHz operating frequency be able to work at 0.651GHz frequency with 65% less area and more accurate results [9][10]. Figure 2 & 4 shows the comparison of return loss & bandwidth of the RMPA alone and with the metamaterial. It has been found that the return loss is reduced by 17dB & the bandwidth is increased by 3MHz of the proposed structure. The Figure 6 & 7 shows the negative value of permittivity & permeability at the operating frequency of 0.651GHz. This proves that the proposed Design of media with a negative refractive index is a Metamaterial Structure.

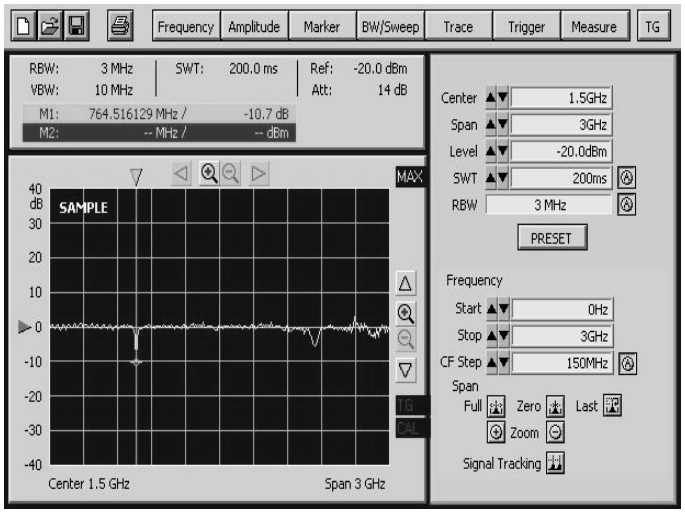


Figure 11: Analyzed result after negative media incorporation showing return loss at 0.76 GHz.

Authors presented a new design methodology in this letter for creating highly miniaturized patch antennas, by adding a single layer that contains a combination of hexagonal and pentagonal like structure at a height of 3.276 on RMPA. The size of the antenna can be reduced significantly without affecting bandwidth with little effort at low cost. The purpose of this work is to produce a small, low cost Antenna that can be used for L band (1-2GHz) applications. An even smaller antenna is possible by this proposed design, but with further miniaturisation comes lacking in radiation efficiency and bandwidth that may prove undesirable.

REFERENCES

[1]. Constantine A. Balanis, Antenna Theory and Design. John Wiley & Sons, Inc., 1997.

[2]. David M. Pozar, "Microwave Engineering", 3rd Edition, John Wiley & Sons, 2004.

[3]. W.L. Stutzman, G.A. Thiele, Antenna Theory and design,John Wiley & Sons, 2nd Ed., New York, 1998.

[4]. Nadar Engheta, Richard W. Ziolkowski, "Metamaterial Physics & Engineering Explorations".

[5]. V. G. Veselago "The electrodynamics of substances with simultaneously negative value ε and μ" Sov. Phys.Uspekekhy.10 (4), 509-514, 1968.

[6]. J.B. Pendry, A.J. Holden, D.J. Robbins, W.J. Stewart,"magnetism from conductors and enhanced nonlinear phenomena" IEEE Trans. Micro Tech. vol.47 no.11, pp.2075-2081, Nov.1999.

[7]. J.B. Pendry, Negative refraction males a prefect lens, Phys Rev Lett, 85 , 3966–396,2000

[8]. R. O. Ouedraogo and E. J. Rothwell, "Metamaterial-inspired patch antenna miniaturization technique," in IEEE Int. Symp. Antennas and Propagation and URSI Radio Science Meeting Dig., 2010, pp. 1–4.

[9]. S. Jahani, J. Rashed-Mohassel, and M. Shahabadi, "Miniaturization of Circular Patch Antennas Using MNG Metamaterials, "IEEE Antennas and Wireless Propagation Letters", vol. 9 (2010): 1194-1196.

[10]. K. L. Wong, "Compact and Broadband Microstrip Antennas". Hoboken, NJ: Wiley, 2002.

[11]. Y. Lee, S. Tse, Y. Hao, and C. G. Parini, "A compact microstrip antenna with improved bandwidth using complementary split-ring resonator (CSRR) loading," in IEEE Int. Symp. Antennas and Propagation and URSI Radio Science Meeting Dig., 2007, pp. 5431–5434.

[12]. R. O. O. and E. J. Rothwell, A. R. Diaz, K. Fuchi, Andrew Temme "Miniaturization of patch antennas using metamaterial inspired technique" IEEE transactions on antennas and propagation, vol. 60, no. 5, may 2012

[13]. F. Bilotti, A. Alu, and L. Vegni, "Design of miniaturized metamaterial patch antennas with negative loading, " IEEE Trans. Antennas Propag., vol. 56, no. 6, pp. 1640–1647, Jun. 2008.

[14]. P.K. Singhal, Bimal Garg "Design and Characterization of Compact Microstrip Patch Antenna Using "Split Ring" Shaped Metamaterial Structure" published in international journal of electrical and computer engineering, Vol.2, No.5, October 2012, pp. 655~662, IJECE.

[15]. H. A. Mazid, M. K. A. Rahim, T. Masri, "Left-handed metamaterial design for microstrip antenna application", IEEE International RF and Microwave conference, 2008.

[16]. D.R. Smith, W.J. Padilla, D.C. Vier, et al, Composite medium with simultaneously negative permeability and permittivity, Phys Rev Lett 84, 4184–4187,May 2000.

[17]. P K Singhal, BimalGarg, NitinAgrawal "A High Gain Rectangular Microstrip Patch Antenna Using "Different C Patterns" Metamaterial Design In L-Band", published in Advanced Computational Technique in Electromagnetics Volume 2012, Article ID acte-00115, 5 pages, ISPACS.

A Reversible Image Steganographic Algorithm Based on Slantlet Transform

Sushil Kumar[1] and S. K. Muttoo[2]

Submitted in November 2012, Accepted in March 2013

Abstract - In this paper we present a reversible image steganography technique based on Slantlet transform (SLT) and using advanced encryption standard (AES) method. The proposed method first encodes the message using two source codes, viz., Huffman codes and a self-synchronizing variable length code known as, T-code. Next, the encoded binary string is encrypted using an improved AES method. The encrypted data so obtained is embedded in the middle and high frequency sub-bands, obtained by applying 2-level of SLT to the cover-image, using thresholding method. The proposed algorithm is compared with the existing techniques based on wavelet transform. The Experimental results show that the proposed algorithm can extract hidden message and recover the original cover image with low distortion. The proposed algorithm offers acceptable imperceptibility, security (two-layer security) and provides robustness against Gaussian and Salt-n-Pepper noise attack.

Index Terms - Reversible Steganography, DWT, SLT, Thresholding scheme, PSNR, AES, Huffman codes, T-codes

1. INTRODUCTION

Data hiding or steganography is the art and science of hiding information into a carrier media (such as text, image, audio or video etc.) so that it conceal the existence of a hidden information and its detection becomes difficult. There are applications in which it is desirable to recover the original cover from the stego-image without any distortion after hidden data extraction. There are many papers on reversible steganography in literature [12, 15, 20-24]. The summary of such algorithms may be seen in the papers [2], [3].

The three basic requirements of steganography algorithm are Imperceptibilty, high embedding payload and security [9, 10, 16]. The organizations such as banking, commerce, diplomacy and medicine, private communications are essential. Security is an important issue in the information technology now-a-days. Modern cryptography provides a variety of mathematical tools for protecting privacy and security that extend far beyond the ancient art of encrypting messages. However, for carrying out confidential communication over public networks, simply concealing the contents of a message using cryptography is found to be inadequate as it can still raise suspicion to eavesdroppers. People have found the solution to this problem in Steganography. The image steganography techniques may be classified into two categories: Reversible techniques in which receiver wish to retain the original message after extracting the hidden message from the stego-image and Irreversible techniques in which the objective of receiver is only in extracting the hidden message from the stego-image. In medical profession and law enforcement fields, it is not only the hiding and recovery of message required perfectly but also the recovery of original image is important for the examination. The authors have used synonyms to Reversible technique as distortionless or lossless technique. Xuan et al. [20-23] have presented distortionless data hiding based integer wavelet transform. Celik et al. [3] have proposed a reversible data hiding method based on the idea of first compressing portion of the signal that are susceptible to embedding distortion and then transmitting it as part of embedded payload. Sushil Kumar and S.K. Muttoo[12] have proposed a distotionless steganographic algorithm based on slantlet transform and shown that it outperforms than the DWT in terms of PSNR. Panda and Meher [13] have shown that Slantlet Transform (SLT) offers superior compression performance compared to the conventional DCT and the DWT based approaches. Ni et al. [12] presented a reversible data hiding algorithm based on histogram shifting with a quite limited embedding payload Tian [17] proposed a high capacity reversible data hiding scheme by using a difference expansion. Xian-ting Zeng et al. [24] have proposed a lossless data hiding scheme by using dynamic reference pixel and multi-layer embedding. This scheme can offer very high embedding capacity and low image degradation.

In this paper, we propose a reversible image steganographic method based on CTT. The proposed scheme can offer high imperceptibily than the existing scheme based on DWT and low image degradation. The use of T-code is a plus point as it provides self-synchronization at decoding stage and a layer of security as receiver will need decoding key (generated at the time encoding) for extracting the original message at decoding stage. There is another layer of security added at embedding scheme by hiding the secret bit randomly, i.e., using random permutation of sub-bands coefficients. Advanced encryption standard (AES) used in the scheme is one of the most powerful techniques of cryptography which can be used as an integral part of steganographic system for better confidentiality and security. Dilbagh singh et al. [4] has proposed private key encryption technique that can be used for data security in modern cryptosystem. Their technique uses the concept of arithmetic coding and can also be clubbed with any of the encryption system that works on floating point numbers.

The rest of the paper is organized as follows: Section 2 presents a review of Slantlet Transform. We introduce briefly the thresholding algorithm applied for embedding in our

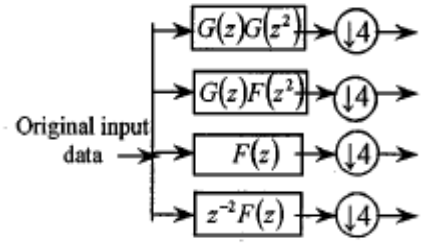
[1]*Rajdhani College, University of Delhi, New Delhi, India*
[2]*Department of Computer Science, University of Delhi, Delhi, India, E-mail ,*[1]*azadsk2000@yahoo.co.in and*
[2]*skmuttoo@cs.du.ac.in*

method in Section3. Section 4 presents the proposed algorithms. The experimental results and their analysis is presented in Section 5. Conclusions and future scope are presented in Section 6.

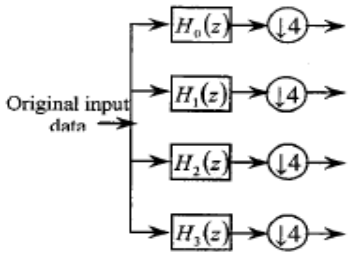
2. SLANTLET TRANSFORM

In image compression, the Wavelet transforms produces much less blocking artifacts than the DCT. They are adopted in JPEG2000. They also perform well in image de-noising. However, 2D wavelet transform is, intrinsically, a tensor-product implementation of the 1D wavelet transform, and it provides local frequency representation of image regions over a range of spatial scales, and it does not represent 2D singularities effectively. Therefore it does not work well in retaining the directional edges in the image, and it is not sufficient in representing the contours not horizontally or vertically.An orthogonal discrete wavelet transform with approxim ation order two, i.e., with two zero moments and improved time localization, known as Slantlet transform (ST), was introduced by Ivan W. Selesnick[14] in 1999. It uses a special case of a class of bases described by B. Alpert et al. [1], the construction of which relies on Gram-Schmidt orthogonalization. It is based on a filterbank structure, implementing in a parallel form, employing different filters for each scale. In DWT, some of these parallel branches employ product of basic filters, shown in figure 1. The Slantlet filter branches, however, do not employ any product form of implementation, as shown in figure 2 and hence ST possesses extra degrees of freedom. Ivan W. Selesnick [14] has shown that due to this property, ST can be implemented employing filters of shorter supports, and yet maintaining the desirable characteristics like orhtogonality and an octave-band characteristics, with two zero moments. For k=2, the iterated filters of Daubechies are of length 10 and 4 whereas in case of SLT they are of length 8 and 4, i.e., 2-scale SLT filterbank has a filter length which is two samples less than that of a 2-scale iterated D_2-filterbank. This difference grows with the number of stages. Though SLT has no tree structure like DWT, it can be efficiently implemented with same order of complexities as of DWT.

Data compression using 2-scale SLT filterbank involves three steps: transformation of input signal using the SLT, thresholding of transformed coefficients and reconstruction of the signal from the thresholded coefficients. G. Panda et al [13] have shown that SLT provides improves time localization than the DCT and DWT.



(a)

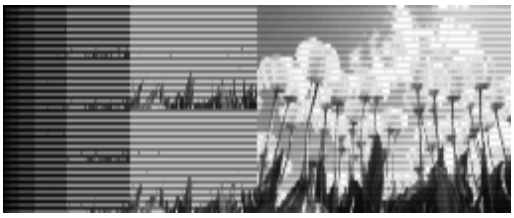


(b)

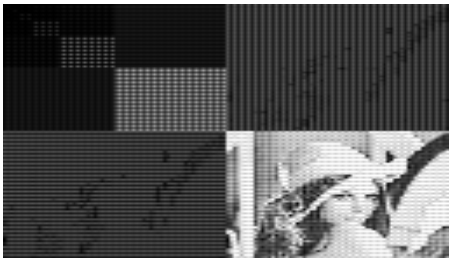
Figure 1: (a) Two-scale iterated filterbank using the DWT (b) Two-scale filterbank structure using the SLT.

Also considering various compression parameters such the percentage of energy retained and the MSE of different PQ signals, it is observed by them that the accuracy of the reconstruction of SLT method is better than that the DCT and DWT, i.e., the SLT based compression technique yields better performance compared to both the DCT and DWT. In the compression scheme using SLT, the data is first applied to two-level filter structures $H_0(z)$, $H_1(z)$, $H_2(z)$, and $H_3(z)$. The output of these filters are down sampled by a factor of 4, which are the transform coefficients of the input data obtained after the convolution operation of the original data with the filter coefficients , as shown in figure 1. The transform coefficients are then thresholded using a suitable parameter. The inverse slantlet transform are performed on these thresholded coefficients to reconstruct the original data.

The figure 2(a) is the 1-level decomposition obtained after applying 1-d slantlet filters to image 'Tulips.jpg' and decomposing into low (L) and high sub-bands(H). The figure 2(b) shows the 2-level decomposition of image lena. bmp when the 1-D slantlet filters are used first on the rows of image and then on the columns, resulting into sub-bands HH, HL, LH and LL respectively.



(a)



(b)

Figure 2. a) 1-level Slantlet image of "Tulips.jpg", and b) 2-level Slantlet image of 'lena.bmp'

Nagaraj B Patil et al [11] have shown that as threshold level increased better compression ratio and PSNR can be achieved for the test data. It has been observed that most of the middle and high frequency coefficients in the HL, LH or, HH subbands obtained from SLT are of low magnitudes. As these bands constitute 75% of all SLT coefficients, the highest payload can be 0.75 bit per pixel (bpp). Table 1 lists the payload of four different images (256x256) under different thresholds "alternately" (unless you really mean something that alternates). For 'Flower.jpg', if threshold T is set to be 8, the payload is 0.645 bpp. It shows that over 86% coefficients in the high frequency subbands are used for data hiding in the Threshold embedding technique.

Image	T=4	T=6	T=8
Lena	0.547	0.58	0.603
Tulips	0.504	0.538	0.566
Flower	0.589	0.623	0.645
Bunkbed	0.50	0.54	0.57

Table 1: Threshold vs payload

3. THRESHOLDING METHOD

Threshold embedding method for the lossless data hiding is given by Xuan et al. [21]. We predefine a threshold value. To embed data into a high frequency coefficient of sub-band HH, LH or HL, the absolute value of the coefficient is compared with T. If the absolute value is less than the threshold, the coefficient is doubles and message bit is added to the LSB. No message bit is embedded, however, the coefficients are modified as follows:

$$x' = \begin{cases} 2*x + b & \text{if } |x| < T \\ x + T & \text{if } x \geq T \\ x - (T-1) & \text{if } x \leq -T \end{cases}$$

where T is the threshold value, b is the message bit, x is the high frequency coefficient and x' is the corresponding modified frequency coefficients.

To recover the original image, each high frequency coefficient can be restored to its original value by applying the following formula:

$$x = \begin{cases} \lfloor x'/2 \rfloor & \text{if } -2T < x' < 2T \\ x' - T & \text{if } x' \geq 2T \\ x' + T - 1 & \text{if } x' \leq -2T + 1 \end{cases}$$

The Figure provides an example to hide the message, s=101010 into a block of 3x3 where T=10.

3	**-7**	**11**
12	**0**	**-1**
8	**-15**	**5**

Embedding
T=10; s=101010

Extraction

6	**-13**	**21**
18	**0**	**-1**
16	**-24**	**11**

4. PROPOSED ALGORITHM

The proposed reversibhle image steganography algorithm embeds data into the first level high frequency subbands of the cover image. Preprocessing is performed prior to data embedding to ensure that no overflow/underflow takes place. The stego-image carrying hidden message is obtained after taking the inverse contourlet transform. Fig. 3 is the flowchart of the proposed embedding data hiding ande Figure 4 is the flowchart for hidden data extraction and original cover image recovery.

The embedding algorithm is summarized as follows:

Algo: Embedding

……

Step1. First obtain the secret data by applying best T-codes as a source encoder to the given input text/message.

Step 2. Modified AES encryption algorithm [25] is applied on the compressed data.

Step3. Apply pre-processing to prevent possible "overflow" during embedding (e.g., replacing the grayscale values 0 to 255 into 15 to 240).

Step4. Consider 8-bit greyscale image and decompose it into 4 sub-bands : one lowpass sub-band and 3 sub-bands for horizontal and vertical directions by applying 2-level SLT, viz., HL,LH and HH

Step5. Embed data in the high horizontal and vertical sub-bands of SLT using thresholding method (taking threshold value=35).

Step6. Obtain the stego-image by taking the inverse SLT of the modified image of step5.

…….

Algo: Extraction

…….

Step 1. Apply CTT to the stego image

Step 2. Extract secret data from the four horizontal and vertical subands of CTT inverse thresholding technique.

Step 3. Improved AES decryption algorithm[21] is applied on the extracted codes to obtain the actual encoded T-codes.

Step 4. Obtain the original message by T-decoding the secret data, with the help of encoding key

Step 5: Recover the original image by removing the hidden message from the stego-image

……

IMAGE	WLT +HUFF +AES	WLT +HUFF +AES (adding Gaussian)	SLT +HUFF +AES	SLT +HUFF +AES (adding Gaussian)
I1	19.922627	19.922627	23.235624	21.450025
I2	18.188314	18.188314	34.095894	32.143663
I3	17.292913	17.292913	26.616492	24.477793
I4	17.454110	17.454110	25.362884	23.197050

Table 2: PSNR values based on Wavelet and SLT using Huffman encoding (secret message = 5000 bits)

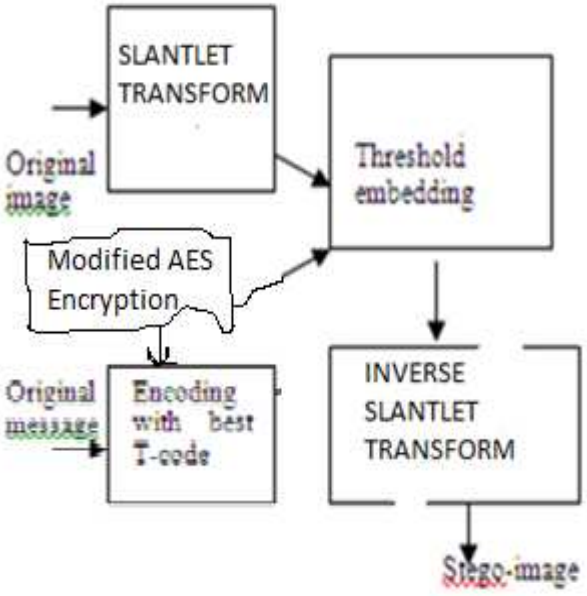


Figure 3: Block diagram of Embedding method

5. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the performance of the proposed data hiding algorithm, we have used 128 x128 and 256 x256 gray scale images. Simulations are done using MATLAB 8.0. We have compared the performance of the proposed steganographic method based on SLT using T-codes as endcoder, improved AES as encryption and reversible thresholding technique as embedding with the corresponding steganograpic method based on Wavelet. We have tested number of images such as standard images and medical images. We have used the metric PSNR for measuring the stego-image quality.

Imperceptibility

The perceptibility measure for the quality of image used is PSNR given by

$$PSNR = 10 \log_{10} (255^2 / MSE)$$

$$MSE=(1/N)^2 \sum\sum (x_{ij} - x'_{ij})^2$$

where x denotes the original pixel value

Table 2 shows the test results for these methods using only Huffman codes as encoder, Table 3 shows test results using only T-codes as encoder, Table 4 shows the results using Huffman codes and improved AES encryption, and Table 5 shows the results using T-codes and modified AES encryption. We have shown the results for the four images: I1:Cameraman.tif, I2: Lena.jpg, I3: Nature.jpg, and I4: Scenery.jpg (see Figure 9).

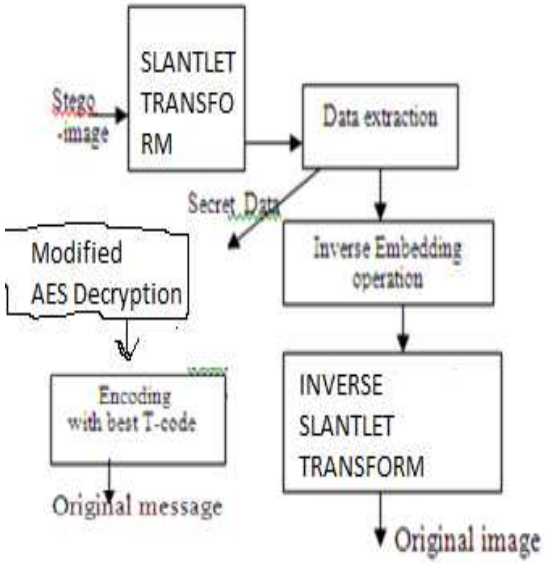


Figure 4: Block diagram of Extraction method

IMAGE	WLT+HUFF	WLT+HUFF (adding Gaussian)	SLT+HUFF	SLT+HUFF (adding Gaussian)
I1	19.921678	19.921678	21.452172	21.452389
I2	18.203956	18.203956	32.140011	32.137198
I3	17.292666	17.292666	24.489907	24.489990
I4	17.453638	17.453638	23.198266	23.200940

Table 3: PSNR values based on Wavelet and SLT using T-code encoding (secret message = 5000 bits)

IMAGE	WLT + TCODE	WLT + TCODE (adding Gaussian)	SLT+ TCODE	SLT+ TCODE (adding Gaussian)
I1	19.276835	18.739734	21.144950	21.146281
I2	16.892371	16.798323	31.866441	31.871396
I3	15.368473	18.578029	24.3046	24.296260
I4	14.086282	9.738723	22.955329	22.951326

Table 4: PSNR values based on Wavelet and SLT using Huffman encoding and AES encry*ption* (secret message = 5000 bits)

Robustness

The figures 5 to 8, we show the bar diagrams for comparison of PSNR values for four images using the proposed algorithm based on slantlet transform, T-codes andf AES method with or without the additon of Gaussian noise (0.01) and compared with the corresponding algorithm for wavelet transform.

Analysis

The results of the PSNR of the proposed method based on SLT is compared with the Wavelet transform and Slantlet transform and are summarized in the table 2 to table 5.

The imperceptibility is found to be better in the SLT based reversible thresholding algorithm than DWT based reversible thresholding method.

IMAGE	WLT+TCODE +AES	WLT +TCODE +AES (adding Gaussian)	SLT+TCODE +AES	SLT+TCODE +AES (adding Gaussian)
I1	18.739734	19.276835	21.143900	21.1446
I2	16.798323	16.892371	31.859676	31.8704
I3	18.578029	15.368473	24.295226	24.2963
I4	9.738723	14.086282	22.952757	22.9471

Table 5: PSNR values based on Wavelet and SLT using T-codes encoding and AES encry*ption* (secret message = 5000 bits)

The algorithm does not need original image for recovering the secret data (It is a blind data hiding scheme). The use of T-codes provides self-synchronization in the decoding stage.

From the above tables it can be seen that SLT along with Huffman compression technique and AES encryption method has slightly better PSNR values than SLT along with T-codes and AES method.

Further, SLT based steganographic method is robust to Gaussian effect (same results have been observed for salt and pepper).

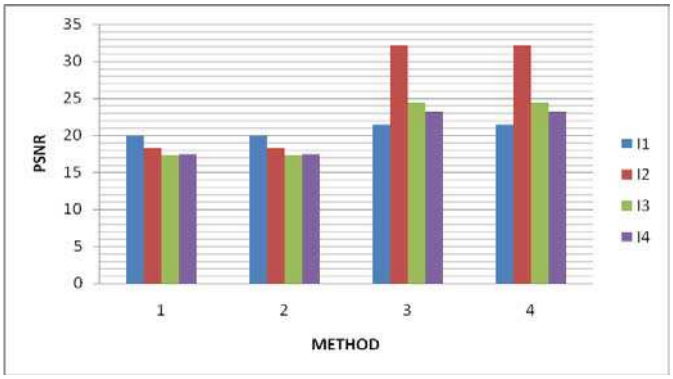


Figure 5: (1) WLT+Huff, (2) WLT+Huff+Gaussian (0.01), (3) SLT+Huff , (4) SLT+Huff+Gaussian

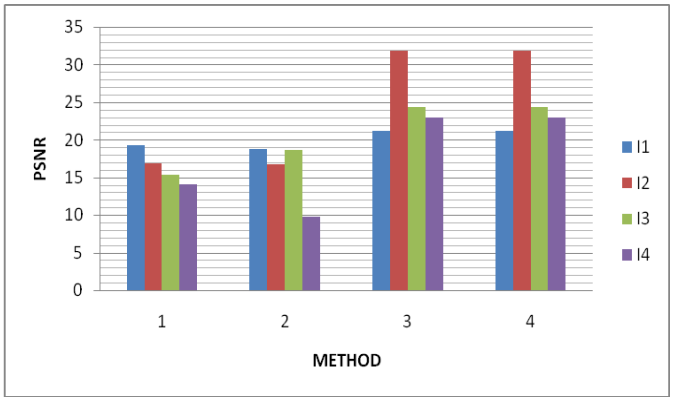


Figure 6: (1) WLT+Tcode, (2) WLT+ Tcode +Gaussian (0.01), (3) SLT+Tcode , (4) SLT+Tcode +Gaussian

6. CONCLUSION AND FUTURE SCOPE

In this paper we have presented

1. a new variable length codes, viz., T-codes for the compression of embedding message.
2. An improved AES for the encryption of the encoded message
3. SLT in place of DWT as they provide better perceptibility and compression.
4. The reversible thresholding technique so that one can recover the original image from the stego-image.

Slantlet transform, which is also a wavelet-like transform and a better candidate for signal compression compared to the DWT based scheme and which can provide better time localization,

Huffman codes have been preferred for data compression by researchers. However, people have been searching for self synchronizing variable length codes since 1970. One of the best self-synchcronization variable length codes which can replace Huffman code is T-code [18-19].We have applied these codes for data compression in the proposed algorithm.

The T-codes are self-synchronizing codes shown to be better than Huffman codes in the decoding process. They also provide a layer of security in the system as one needs encoding key to encode the secret message obtained from the extraction process.

The use of encryption in steganography can lead to ‘security in depth’. To protect the confidential data from unauthorized access, an advanced encryption standard (AES) has been suggested by the researchers [5]. AES algorithm is a very secure technique for cryptography and the techniques which use frequency domain are considered highly secured for system for the combination of steganography.

The reversible threshold embedding technique is used for embedding the secret message in the sub-bands of transform image obtained from the cover object by applying 2-level of SLT and results are compared with the data hiding techniques based on wavelet (biorthogonal cdf9/7) transform.

Security
The integration of Compression technique (T-codes) and cryptography technique (Modified AES) with Steganography use three keys – encoding key, encrypted key and threshold value, making the present algorithm a highly secured method.

Robustness
The proposed method provides not only acceptable image quality but also has almost no distortion in the stego-image after adding Gaussian noise or Salt and Pepper noise. The use of SLT has shown better results than DWT in terms of image metric 'PSNR' and robustness.

Recovery
There is no artifact obtained in the stego-image and the original image is recovered with low image degradation from the stego-image.

Embedding Payload
The embedded payload in the proposed embedding technique is same as in case of the DWT techniques.

REFERENCES

[1]. Alpert B., Coifman G.R., and Rokhlin V., "Wavelet-like bases for the fast solution of second kind integral equations", SIAM J. Sci. Comput., 14: 159-184, 1993, http:///dx.doi.org/10.1137.0914010

[2]. Awrengjeb M., "An Overview of reversible Data Hiding", ICCIT 2003, Jahangir Nagar University, Bangladesh, Dec. 19-21, pp. 75-79, 2003

[3]. Celik M., Sharma G., Taekalp A.M., Saber E., "Reversible data hiding", in Proceeding of the International Conference on Image Processing 2002, Rochester, NY, September, 2002.

[4]. Singh Dilbagh and Singh Ajay, "A secure private key encryption technique for data security in modern cryptosystem", BIJIT - BVICAM's International Journal of Information Technology, Vol. 2, No. 2, 2011

[5]. Domenico Daniele Bloisi, Luca Iocchi,"Image based Steganography and cryptography", Computer Vision theory and applications volume 1, pp. 127-134, 2007.

[6]. Maitra M. and Chatterjee A., "A Slantlet transform based intelligent system for magnetic resonance brain image classification", Biomedical Siganl Processing and Conrol, 1 , pp. 299-306, 2006

[7]. Muttoo S.K. and Sushil Kumar,"Data Hiding in JPEG images", BIJIT - BVICAM's International Journal of Information Technology, Vol. 1, No.1, Jan.-July, 2009.

[8]. Muttoo S.K. and Sushil Kumar, "A mulltilayred secure, robust and high capacity image steganographic algorithm", World of Computer Science and Information Technology Journal (*WCSIT*), ISSN 2221-0741, vol. XXX, No. XXX, 2011

[9]. Muttoo S.K. and Sushil Kumar," Robust Source coding steganographic technique using Wavelet Transforms", BIJIT - BVICAM's International Journal of Information Technology Vol. 1, No. 2, July – December, New Delhi, 2009

[10]. Muttoo S.K. and Sushil Kumar, "Robust source coding watermarking technique based on magnitude DFT decomposition", BIJIT - BVICAM's International Journal of Information Technology, Vol. 4, No. 2, 2011

[11]. Nagaraj B. Patil, V.M. Viswanatha, Dr. Sanjay Pandey M.B., "Slant Transformation as a tool for pre-processing in image processing", Int. J. of Scientific 7 Engineering Research, Vol. 2, Issue 4, April-2011.

[12]. NI Z., Shi Y.Q., Ansari N., Su Wei, Sun Q. and Lin X., "Robust Lossless Image Data Hiding", IEEE International Conference on Multimedia and Expo(ICME), 2199-2202, 2004

[13]. Panda G. and Meher S.K., "An efficient approach to signal compression using slantlet transform", IETE Journal of Research, Vol. 46, No. 5, September, pp. 299-307, 2000.

[14]. Selesnick Ivan W., "The Slantlet Transform", IEEE transactions on signal processing, Vol. 47, No. 5, May, pp. 1304-1312,1998.

[15]. Sushil Kumar and S.K. Muttoo, "Distortionless Data Hiding based on Slantlet Transform", Proceeding of the first International conference on Multimedia Information Networking & Security (Mines, 2009), Wuhan, China, Nov. 17- 20, Vol. 1, pp. 48-52, IEEE Computer Society Press, 2009

[16]. Sushil Kumar and S.K. Muttoo, "An Overview on Wavelet-like Transforms in Image Data Hiding", Proceeding of the National Conference on computing for Nation development, 2010, New Delhi, Feb. 25-.26, pp. 263-268

[17]. Tian J., "High capacity reversible data embedding and content authentication", IEEE International Conference on Acoustic, Speech, and Signal Processing, April 6-10, vol. 3, pp. 517-520, 2003.

[18]. Titchener, M. R., "Generalised T-codes: extended construction algorithm for self- synchronization codes", IEE Proc. Commun., Vol. 143, No.3, pp. 122-128, 2006.

[19]. Ulrich G., "Robust Source Coding with Generalised T-codes", a thesis submitted in the University of Auckland, 1998.

[20]. Xian-ting Zang, Zhuoh Li and Ling-di Ping, "Reversible data hiding scheme using reference pixel and multi-layer embedding", Int. J. of Electronics and Commun. (AEU), 66, 532-539, 2012

[21]. Xuan G., Zhu J., Chen J., Shi Y.Q., Ni Z, and Su W., "Distortionless data hiding based on integer wavelet transform", IEEE Electronics Letters, Dec., pp. 1646-1648, 2002.

[22]. Xuan G.,.Shi Y.Q., Yang C., Zhang Y., Zou D. and Chai P., "Lossless Data Hiding using integer wavelet transform, and threshold embedding technique", in

Proceeding of IEEE International Workshop on Mutimedia Signal Processing, Marriott Beach Resort ST. Thomas, US Virgin Islands, Dec. 9-11, 2002.

[23]. Xuan G., Yang C., Zhen Y., Shi Y.Q., Ni Z., " Reversible data hiding based on wavelet spread spectrum", IEEE 6th Workshop on Multimedia Signal Processing, 211-214, 2004.

[24]. Yu L. and Sun S., "Slantlet transform based image fingerprints", Communication, Netwirk and information security, CNIS, 2000.

[25]. M. Zeghid, M. Machhout, L.Khriji, A. Baganne and R. Tourki,"A Modified AES Based Algorithm For Image Encryption", World Academy Of Science, Engineering and Technology 27, 2007

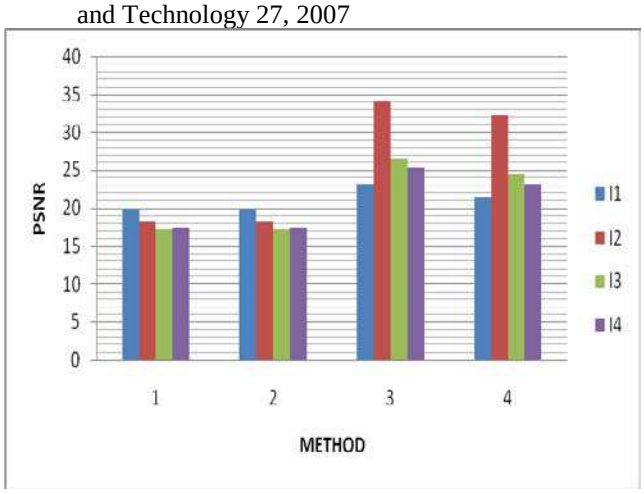


Figure 7: (1) WLT+AES+Huff, (2) WLT+AES+ Huff+Gaussian (0.01), (3) SLT+AES+Huff , (4) SLT+AES+Huff+Gaussian

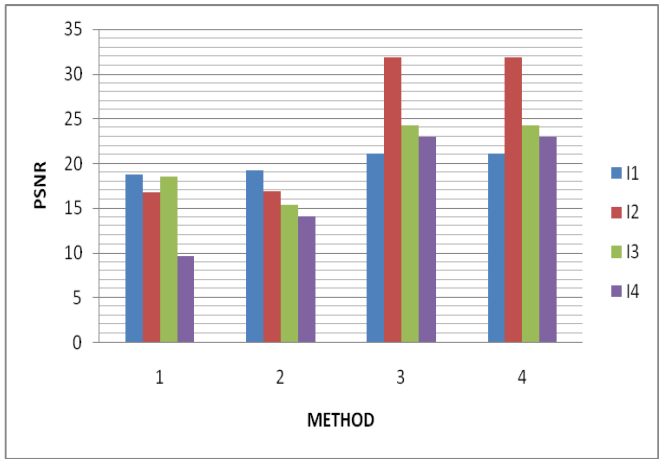


Figure 8: (1) WLT+AES+Tcode, (2) WLT+AES+ Tcode +Gaussian (0.01), (3) SLT+AES+ Tcode , (4) SLT+AES+ Tcode +Gaussian

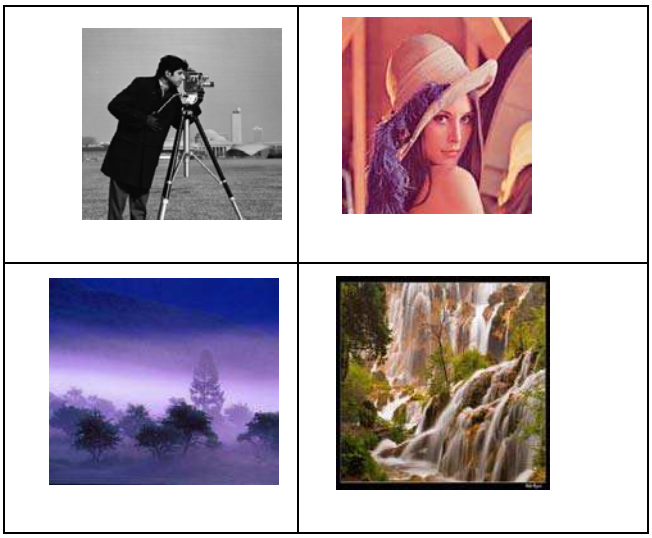


Figure 9: Cover images I1, I2, I3 and I4

Performance Analysis of Massively Parallel Architectures

Z. A. Khan[1], J. Siddiqui[2] and **A. Samad[3]**

Submitted in October 2012, Accepted in April 2013

Abstract - Cube based networks have received much attention over the past decade since they offer a rich interconnected structure with a number of desirable properties such as low diameter, high bisection width, lesser complexity and Cost. Among them the hypercube architecture is widely used network for parallel computer system due to its low diameter. The major drawback of hypercube based architectures is the difficulty of its VLSI layout. Several variations of hypercube have also been reported which are designed by considering a specific topological property. Nevertheless, no particular topology claims to have better performance with all the desirable topological properties. In this paper the performance analysis of various interconnection networks is presented. The performance is compared by considering cube type architectures as well as linear type architectures on different parameters such as degree, diameter, bisection width, scalability and cost etc. The Analysis indicates that cube based architectures have a rich interconnected structure with high cost and complexity. On the other hand linear type architectures are scalable, simpler and better in terms of cost and complexity. The comparative study suggests the various aspects to the design of new multiprocessor architectures.

Index Terms - Interconnection network, Performance evaluation, Topological properties, Parallel system, Cube Architectures

1. INTRODUCTION

systems interconnection networks play an important role in the overall performance of the system. Deciding the appropriate network is an important issue in the design of parallel and distributed systems. In general, determining the optimal network to implement any parallel application does not have a known theoretical solution. There are different ways to determine efficient topologies that trade-off high level performance issues against various implementation constraints [1]. A Topology is evaluated in terms of a number of performance parameters such as degree, diameter, bisection width and cost. Several researchers have developed various architectures which are considered better in terms of particular parameters.

[1,2]Dept. of Computer Science, Aligarh Muslim University, Aligarh (India)-202002
[3]University Women's Polytechnic, Aligarh Muslim University, Aligarh (India)-202002
E-Mail: [1]khanzaki05@gmail.com,
[2]jamshed_faiza@rediffmail.com and
[3]abdussamadamu@gmail.com

Some variations focus on the reduction of the diameter [10] [18], some of them focused on the design of simple routing and communication algorithm [4]. Scalability is also an important issue to evaluate the performance of interconnection networks. However, it can't be clearly mentioned that which interconnection network is working better by considering all the parameters. In terms of complexity interconnection networks may be classified into two major categories. The first is cube based architectures which posses a rich interconnection topology. The Binary hypercube or n-cube has been widely used interconnection network in the design of parallel systems [12]. Several variations of hypercube architecture are reported in the literatures some examples are –folded hypercube (FDC), metacube (MC), folded metacube (FMC) and folded dualcube (FDC) etc. [8] [7] [12] [13] [15] [11] . The major drawback in such networks is the increase in the number of communication links for each node and the increase in the total number of nodes in the system which ultimately enhances the complexity of such interconnection networks [19] [20]. Therefore, there is a need to carry out the performance analysis of various interconnection networks by considering their topological properties.

The second class of the network is linearly extensible networks such as linear array, ring, linearly extensible tree and linearly extensible cube etc [10] [16]. The complexity of these networks is lesser as they do not have exponential expansion. Besides the scalability, other parameters to evaluate the performance of such networks are degree, number of nodes, diameter, bisection width and fault tolerance. The main purpose of this paper is to study and analyse the various multiprocessor networks along with their properties to help in the design of a new interconnection architecture. Selection of a better interconnection network may have several applications with lesser complexities and improved power-efficiency. One such modern application is network on chip (NoC) paradigm where different cores are embedded with appropriate connectivity. Some examples may include mesh, torus, star, etc. [1] [9].

In this paper, the study of five cubes based architectures as well as several linear extensible architectures are carried out. Section 2 describes, the various parameters used to make the performance analysis. Various parameters used to compare the performance of cube based architectures and their characteristic is discussed in section 3. Similarly, the comparative analysis of linearly extensible architectures is carried out in section 4. A comparative study of both the type of architectures is made in section 5 and finally concluded the paper in section 6.

2. PERFORMANCE PARAMETERS

The need for architectural performance evaluation exists from design phase to its installation. The various parameters decide the design alternatives and gives a criterion of selection known as cost performance trade off [6] [3]. In general, the performance of various architectures is measured by the following parameters.

A. Degree (d)

It is connectivity among different nodes in a network. The connectivity of the nodes determines the complexity of the network. The greater number of links in the network means greater is the complexity.

B. Diameter (D)

It is defined as the maximum shortest path between the source and destination node. The path length is measured by the number of links traversed. This virtue is important in determining the distance involved in communication and hence the performance of parallel systems.

C. Bisection width (B)

The bisection width of a network is the minimum number of edges whose removal will result in two distinct sub networks. Greater bisection width is better for a network to be fault tolerant.

D. Cost (C)

It is defined as the product of the diameter and the degree of the node for the asymmetric network. (i.e. Cost = D*d). This factor is widely used in performance evaluation.

E. Extensibility

This is the virtue which facilitates large sized system out of small ones with minimum changes in the configuration of the nodes. It is the smallest increment by which the system can be expanded in a useful way.

In the Present work the above parameters are compared for different types of multiprocessor architectures. The values are computed based on a certain mathematical formula designed for specific topology.

3. CUBE BASED ARCHITECTURES

A. Hypercube

The Binary hypercube or n-cube has been one of the most popular interconnection networks having logarithm diameter [12]. Each node in this network is connected through bidirectional asynchronous point-to-point communication link to other nodes. The major drawback of the hypercube is the increase in the number of communication links for each node with the increase in the total number of nodes in the system[17]. The hypercube has a high bisection width $b=2^{n-1}$ and has good capability of fault tolerance.

B. Folded Hypercube

The folded hypercube (FHC) is a standard hypercube with some extra links established between its nodes [2]. A folded hypercube of dimension n is called FHC (n). The FHC (n) is constructed from a standard hypercube by connecting each node to the unique node that is farthest from it. The FHC (n) is a regular network of node connectivity (n+1) and the hypercube of degree 3 is converted to FHC network as show in Figure 1. The diameter of an FHC (n) is (n/2) and bisection width is $2^n/4$.

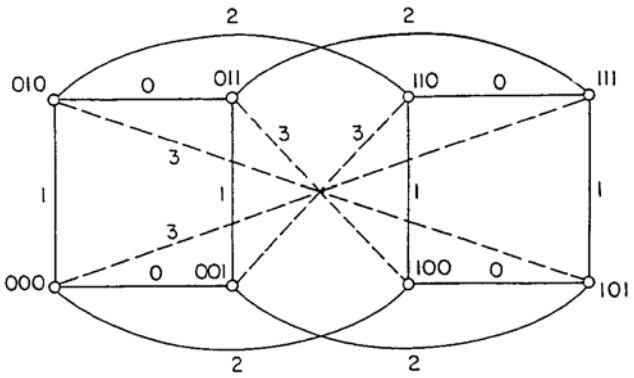


Figure 1: Folded Hypercube FHC (3)

C. Metacube

The metacube (MC) is an interconnection network for a very large parallel computer. In this network, the number of nodes is much larger than hypercube with a small number of links per node [4] [14]. The metacube network shares many desired virtues of the hypercube such as small diameter. The metacube (MC) network includes the dual-cube as a special case. The MC network has two level cube structure a high-level cube (classes) represented by the k- dimension and low- level cube (cluster) represented by m-dimension. An MC (k, m) network can connect 2^{k+m2k} nodes with (k+m) links per node. The degree is $m+k= (n-k)/2^k+k$ and the bisection width of an MC (k, m) is $2^{m2k}/2$.

D. Folded Metacube

The folded metacube is an interconnection topology which inherits some of the useful properties of the metacube and folded hypercube (FHC) [5]. The folded metacube is graph G (V, E) as show in Figure 2. Where V represents a set of vertices and E represents a set of links. The graph is a modified of metacube. The diameter of folded metacube is 2(m+k)-1 and the Bisection width of G is $2^{m2k}/2 + 2^{m2k+k-2}$.

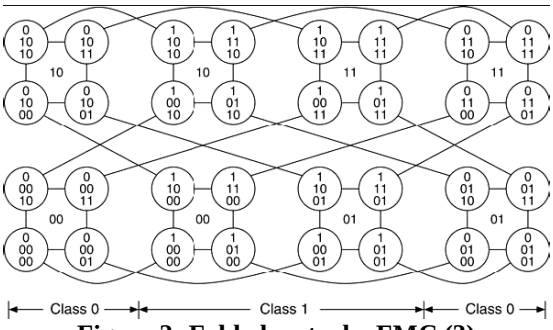


Figure 2: Folded metcube FMC (3)

E. Folded Dualcube

The Folded dualcube (FDC) is a cube based topology which inherits some of the useful properties of the dualcube [8] and the folded hypercube (FHC) [2]. The folded dualcube, which is constructed by connecting each node farthest from it and is show in Figure 3.

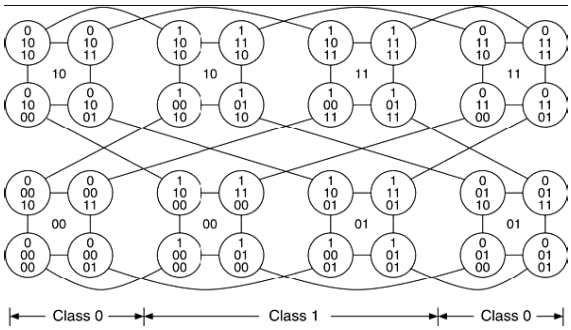


Figure 3: Folded dualcube FDC (3)

The nodes connectivity of folded dualcube is (n+3)/2, the diameter is n-1 and having bisection width is $2^n/2$ [5]. The Various parameters of cube based architectures along with the topological properties are summarized in Table 1.

Type	Nodes	Degree (d)	Diameter (D)	B.W	Cost	Extensibility
HC	2^n	n	n	2^{n-1}	n^2	Exponential
FHC	2^n	n+1	n/2	2^n	n/2*n+1	Exponential
MC	2^n	$(n-k)/2^k+k$	2^{k+1}	$2^{2k}/2$	$(n-k)/2^k+k*2^{k+1}$	Exponential
FMC	2^n	(n+1)	2n-1	$2^{2k}/2+2^{2k}+k-2$	(n+1) *(2n-1)	Exponential
FDC	2^{2n-1}	(n+3)/2	n-1	$2^n/2$	(n+3)/2 * (n-1)	Exponential

Table 1: Various parameters of Cube based Architectures

4. LINEAR EXTENSIBLE ARCHITECTURES

A. Linear Array

It is one dimensional network having the simplest topology with n-nodes having N-1 communication links. The internal nodes have degree 2 and the termination nodes have degree1. The diameter is N-1, which is long for large N and the bisection width is 1. It is asymmetric network.

B. Binary tree

The binary tree is scable architecture with a constant node degree and constant bisection width. In general, an n-level, complexity balanced binary tree should have $N=2^n-1$ nodes. The maximum node degree is 3 and the diameter is 2(n-1).

C. Linearly Extensible Tree

A binary type network topology has been reported [10] shown in Figure 4. The Linearly Extensible Tree (LET) architecture exhibits better connectivity, lesser number of nodes over cube based networks. The LET network has low diameter, hence reduce the average path-length traveled by all message and contains a constant degree per node. The LET network grows linearly in a binary tree like shape. In a binary tree the number of nodes at level n is 2n whereas in LET network the number is (n+1).

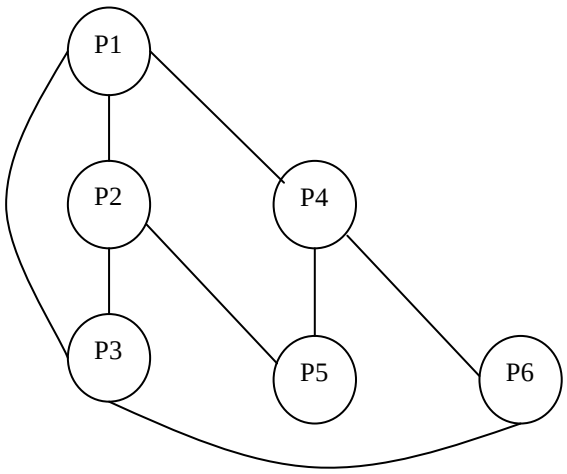


Figure 4. Linearly Extensible Tree (LET) network

D. Linearly Extensible Cube

The Linearly Extensible Cube (LEC) network grows linearly and posses some of the desirable topological properties such as small diameter [10], high connecting constant node degree with high scalability. It has a constant expansion of only two processors at each level of the extension while preserving all the desirable topological properties. The LEC network can maintain a constant node degree regardless of the increase in size (i.e. number of nodes) in a network.

The number of nodes in LEC network is 2*n for n>0 where the number of nodes in the hypercube is 2^n. The diameter of network is $\lfloor N \rfloor$. It has a constant node degree 4. The LEC has a bisection width equal to N**,** as show in Figure 5**.**

Figure 5: Linearly Extensible Cube (LEC) network

E. Ring

A ring is obtained by connecting the two terminal nodes of a linear array with one extra link. A ring network can be uni-or bidirectional and it is symmetric with a constant. It has a

constant node degree of d=2, the diameter is $\lfloor N/2 \rfloor$ for a bidirectional ring and N for unidirectional ring. A ring network has a constant width 2. The different performance parameters of Linearly Extensible Architectures are summarized in Table2.

Type	Nodes	Degree (d)	Diameter (D)	B.W	Cost	Extensibility
LET	$N=\sum_{k=1}^{n} k$	4	$\sqrt{N}$	2log(n-2)	$4\sqrt{N}$	Linear
LEC	N=2*n	4	$\lfloor N \rfloor$	N	$4 \lfloor N \rfloor$	Linear
L.A	N	2	N-1	1	2(N-1)	Linear
Ring	N	2	$\lfloor N/2 \rfloor$	2	$2 \lfloor N/2 \rfloor$	Linear
B.T	$N=2^n-1$	3	2(n-1)	1	6(n-1)	Linear

Table 2: Various parameters of Linearly Extensible Architectures.

5. COMPRATIVE STUDY OF VARIOUS ARCHITECTURE

For multiprocessor network parameters such as diameter, degree, bisection width, cost regularity and symmetry are crucial and determine the performance of the network to compare the performance. We proceed to consider the three important parameters namely, number of processors, diameter and cost. The curves are plotted for each of the parameters for both the class of interconnection networks. Figure 6 shows the trained of increasing number of processors for each level of the extension. It is observed that all the linearly extensible architectures except binary tree have lesser number of processors. Therefore, the complexity of linearly extensible architectures is lesser, when they are expanded on higher level. Having lesser number of processors to implement a parallel algorithm is always economical. On the other hand the cube based architectures have exponential expansions which make the network highly complex. The Figure 6 also shows that among linearly extensible architectures, the LEC network produces better results.

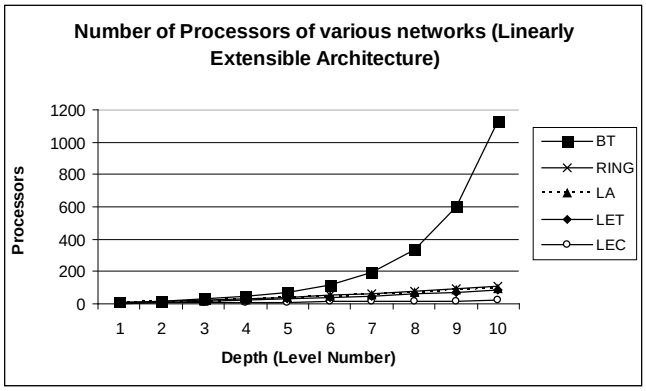


Figure 6: Performance of level extensible architectures

The second parameter when analyzing the performance of both the type of architectures is diameter. To analysis the diameter of various networks the curves are plotted and show in Figure 7 and 8. The study of the results in both the curves shows that the results in both the types of network are comparable. Among

cube based architectures, folded hypercube architectures has lesser diameter as compare to other cube based architectures (Figure 7). When comparing the results of linearly extensible architectures the LEC networks has lesser diameter as compare to other similar architectures.

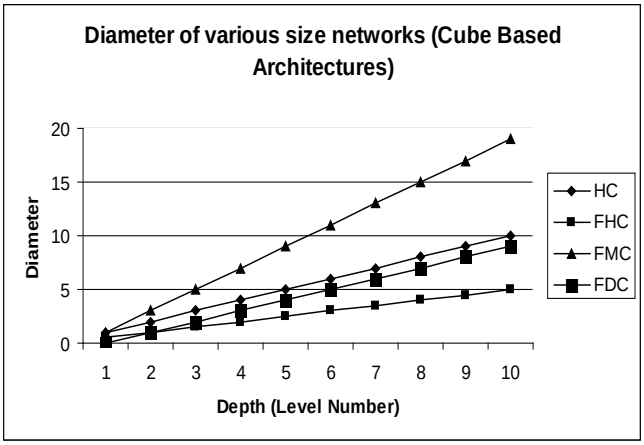


Figure 7: Performance of Cube based architectures

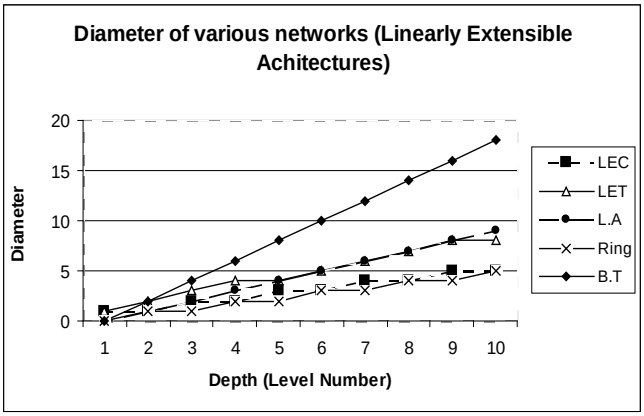


Figure 8: Performance of linearly extensible architectures

The main parameter in terms of evaluating the performance is cost which is defined as the product of the degree and the diameter. Figure 9 and 10 depicts the patterns of the cost analysis of both the class of networks. In cube based network FHC is having lesser cost at greater level as compare to other similar cubical architectures (Figure 9). Similarly, when comparing the cost of linearly extensible architectures, Figure 9 shows that LET is having lesser cost in comparing to other linear types of architectures.

To clearly draw the conclusion, the cost analysis of those architectures is carried out which are giving better results in their respective categories. Therefore, when comparing the cost of FHC and LET, it is observed that LET network has lesser cost at higher level as compare to FHC However, the results are comparable.

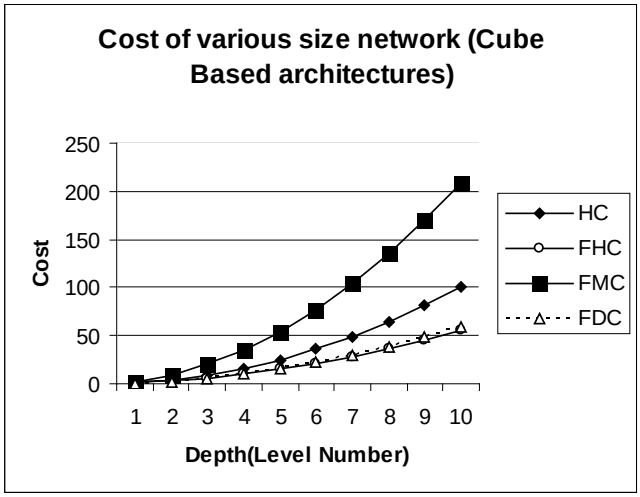


Figure 9: Performance of Cube based architectures

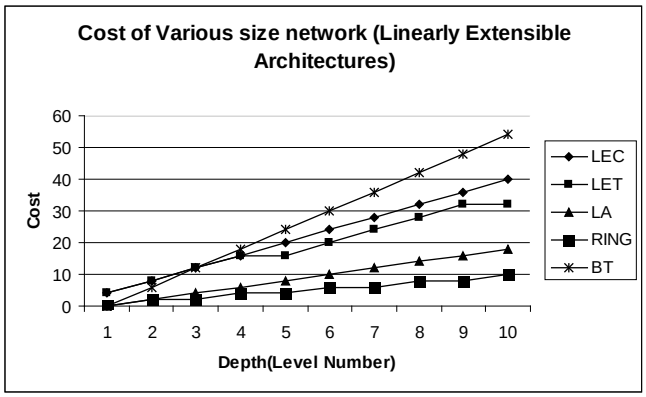


Figure 10: Performance of linearly extensible architectures

The bisection width is also an important parameter for measuring the performance of multiprocessor architectures. The bisection width in cube type architectures is of exponential value. In case of linearly extensible architectures the bisection width is either constant or increases linearly with the increase in number of processors. The linear increment is not desirable, as such, connection at higher level of architectures do not seem to reflect the practical fault tolerance capability of the network.

6. CONCLUSION AND FUTURE SCOPE

In this paper the performance of various multiprocessor architectures are analyzed by considering their topological properties. The comparative study of cube based as well as linearly extensible architectures is made. In cube based networks, it is evaluated that the FHC is giving better performance in terms of diameter and cost. However, all the cube based architectures have exponential expansion which increases the complexity of the system. If we limit the number of processors in FHC it can be considered as best multiprocessor network with high degree of fault tolerance. There is a great scope to modify this network so that it can have approximately all the desirable topological properties with lesser number of processors. As far as linearly extensible architectures are considered they are less complex and easily extensible. However, the common drawback is that they are having low bisection width, which is not a desirable property to make the network fault tolerant.

The important issue in the design of multiprocessor systems is how to cope with the problem of an adequate design of the interconnection network in order to achieve the desired performance at low cost. The choice of the interconnection network may affect several characteristics of the system such as node complexity, scalability and cost etc. The present study is carried out on the basis of several characteristics of various multiprocessor interconnection networks. There have been more work related to design of appropriate multiprocessor network; however no one claims a particular design which entrenched all the desirable properties. The present study gives more scope to design high performance interconnection network that can be used in the design of multiprocessor server.

REFERENCES

[1]. Shafi Patel, Parag Parandkar, Sumant Katiyal and Ankit Agarwal, "Exploring Alternative Topologies for Network-on-Chip Architectures," BIJIT - BVICAM's International Journal of Information Technology - 2011 Vol. 3 No. 2, ISSN 0973 – 5658

[2]. Amway, A. E. and Latifi, S, "Properties and performance of folded hypercubes," IEEE Transactions on Parallel and Distributed Systems, Vol. 2, 1991, pp. 31– 42.

[3]. Harvendra Kumar and A.K. Verma, "Comparative Study of Distributed Computing Paradigms," BIJIT - BVICAM's International Journal of Information Technology – 2009 Vol. 1 No. 2, ISSN 0973 – 5658.

[4]. Yamin Li, Shietung Peng and Wanming Chu, "Metacube: A New Interconnection Network for Large Parallel System," ACSAC02, Australian Computer Science Communications, Vol. 24, No. 4, 2001, pp.29-36.

[5]. N. Adhikari and C.R. Tripathy, "Folded Metacube: An Efficient Large scale Parallel network," 04/2009; DOI:10.1109/IADCC.2009.4809200 in Proceedings of advanced Computing conference, 2009, IACC 2009, IEEE International.

[6]. Neeraj Kumar, "Simulation Study for Performance and Prediction of Parallel Computers," BIJIT - BVICAM's International Journal of Information Technology - 2012, Vol. 4 No. 2, ISSN 0973 – 5658.

[7]. N. Adhikari and C.R. Tripathy, "Folded Dualcube: A New interconnection for Parallel Systems," Proceedings

of 11th Int. Conf. on Information Technology, 2008, 17-18 Dec, pp.-75-78, IEEE, Comp. Society.

[8]. Yamin Li and Shietung Peng, "Dualcube: A New Interconnection Network for High-Performance computer Clusters," International computer symposium, workshop on Computer architecture. December 6-8, 200, ChiaYi, Taiwan.

[9]. Parag Parandkar, Jayesh Kumar Dalal and Sumant Katiyal, "Performance Comparison of XY, OE and DY ad Routing Algorithm by Load Variation Analysis of 2-Dimensional Mesh Topology Based Network-on-Chip," BIJIT - BVICAM's International Journal of Information Technology -2012 Vol. 4 No. 1 ISSN 0973 – 5658

[10]. Samad A, Rafiq M. Q. and Farooq O., "*LEC: An Efficient Scalable Parallel Interconnection Network*," in proceedings of International Conference on Emerging Trends in Computer Science, Communication and Information Technology (CSCIT2010) Nanded, India.

[11]. Vinay Kumar, "Restricted Backtracked Algorithm For Hamiltonian Circuit in Undirected Graph," BIJIT - BVICAM's International Journal of Information Technology – 2010 Vol. 2 No. 2 ISSN 0973 – 5658.

[12]. Y. Saad and M.H. Schultz, "Topological properties of hypercubes," IEEE Trans. Computer. Vol.37, No. 7, 1988, pp.867–872.

[13]. N. Adhikari and C.R. Tripathy, "Folded Crossed cube: A New interconnection for Parallel Systems," International Journal of computer Application (0975-8887) Volume 4, p43. July 2010.

[14]. Yamin Li, Shietung Peng and Wanming Chu, "Efficent Collective Communications in Dual-cube," The journal of super computing, 28, pp.71-90, 2004.

[15]. Y. Q. Zhang, "Folded-crossed hypercube: A CompleteInterconnection Network," Journal of SystemArchitecture, vol. 47, 2002, pp. 917-922, Elsevier Science.

[16]. Nishant Doshi, Tarun sureja, Bhavesh akbari, Hiren Savaliya and Viraj Daxini, "width of Binary Tree,"International journal of computer Applications *0975 – 8887)* Volume*9– No.2, November 2010.*

[17]. S.R. Deshpande and R.M. Jenevein, "Scalability of a binary tree on a hypercube," in proc. Int. Conf. Parallel Processing, 1986, pp. 661 -668.

[18]. L.N. Bhuyan and D.P. Agrawal, "Performance of Multiprocessor Interconnection Network," IEEE Computer,1989 .

[19]. Y.C. Liu, J.F. Fang and C.C. Wu, "On Extensibilities Of Interconnection Networks," IE.(I) Journal-CP,PP.13-16, 2004.

[20]. S. P. Mohanty, B. N. B. Ray, S. N. Patro and A. R. Tripathy, "Topological Properties Of a New Fault Tolerant Interconnection Network for Parallel Computer," icit, pp. 36-40, 2008 International Conference on Information Technology, 2008.

On the Importance of Ensembles of Classifiers

A. K. Saxena

Submitted in November 2012, Accepted in May 2013

Abstract - In this paper, a recent yet powerful technique for classification of datasets is presented. The paper contributes to highlight the importance of an ensemble approach over individual classifiers to achieve better classification accuracy of a classifier. In this paper, given dataset is divided into a number of parts to constitute an ensemble. The ensemble combines these classifiers. An unknown data pattern is tested on the ensemble. Using bagging, majority of voting technique, the performance of ensemble is determined on different sections of datasets. In the paper, six bench mark datasets are used for investigation. Each dataset is trained with 80%, 60% and 50% of the data patterns for classification. The number of classifiers in an ensemble for each data set is changed to 5,7 and 9. As a typical case, k-nearest neighbor (k-NN) classifiers are used with the values of k varying to 1,3 and 5. The classification accuracies of individual classifiers and those of ensembles are computed at each case. After extensive experiments of proposed scheme, by taking random shuffling and selection of data patterns for training and testing, it is observed that in every case, the classification accuracy obtained by ensemble is higher than that obtained by individual classifier.

Index Terms - Classification, Ensemble of classifies, bagging, k-nn classifier.

1. INTRODUCTION

There have been a significant number of research activities in the area of data analysis. The size of database keeps on increasing with useful or redundant data. The task of analysis of the data becomes complex due to presence of these redundant, mostly unwanted pieces of data, commonly called features in a formatted dataset. The role of a classifier is to divide a dataset on the basis of labels or classes of its patterns. In addition to classifying data patterns into different classes, it is also expected from a classifier to predict the label (or more often termed as class) of an unknown pattern, called test pattern. Classification has become a vital component of the study of pattern recognition [1]. Due to the huge amount of data piled up every moment on disks, web spaces and other storage devices, techniques like data mining [2,3], have become quite relevant. Classification is an important step of data mining. Classification is one of the core challenging tasks [4] in mining [5], pattern recognition [1], bioinformatics [6]. The goal of classification [7,8] is to assign a new entity into a class from a pre-specified set of classes.

Dept of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chattisgarh, India, E-Mail: amitsaxena65@rediffmail.com

A classifier needs to be trained before it can be set ready for predicting the class of unknown patterns. The learning of classifier can be made in two manners viz. supervised and unsupervised. In case of supervised learning, the class of every pattern is known in advance at the time of training. In unsupervised learning, class of the training pattern is not given. Commonly, the classifications are based on classification models (classifiers) that are induced from an exemplary set of pre-classified patterns. Alternatively, the classification utilizes knowledge that is supplied by an expert in the application domain. In a typical supervised learning setting, a set of instances also referred to as a training set is given. The labels of the instances in the training set are known and the goal is to construct a model in order to label new instances. An algorithm which constructs the model is called *inducer* and an instance of an inducer for a specific training set is called a *classifier*. There are several well established classifiers such as Fisher's Linear discriminant analysis (LDA) [25], naive Bayes classifier[26], support vector machines, SVM [27], k-Nearest neighbor [28], Neural Networks [29.], fuzzy [30, 40.]. In many examples, idea behind the construction of an ensemble is to combine the classifiers after a weak or non perfect training of individual classifiers. The ensemble so obtained outperforms every individual classier. In fact, human being tends to seek several opinions before making any important decision. Before buying very costly items or taking critical medical decisions, it is a common practice to weight the individual opinions, and combine them to reach to a final decision [9]. Recently, Mikel Galaretal[10] reported that class distribution, i.e., the proportion of instances belonging to each class in a data-set, plays a key role in classification. Sometimes imbalanced data-sets problem occurs when one class, usually the one that refers to the concept of interest (positive or minority class), is under-represented in the data-set; in other words, the number of negative (majority) instances outnumbers the amount of positive class instances [11]. The primary benefit of using ensemble systems is the reduction of variance and increase in confidence of the decision. Due to many random variations in a given classifier model (different training data, different initialization, etc.), the decision obtained by any given classifier may vary substantially from one training trial to another—even if the model structure is kept constant. Then, combining the outputs of several such classifiers by, for example, averaging the output decisions, can reduce the risk of an unfortunate selection of a poorly performing classifier. Another use of ensemble systems includes splitting large datasets into smaller and logical partitions, each used to train a separate classifier. This can be more efficient than using a single model to describe the entire data. The opposite problem,

having too little data, can also be handled using ensemble systems, and this is where bootstrap-based ideas start surfacing: generate multiple classifiers, each trained on a different subset of the data, obtained through bootstrap re-sampling. While the history of ensemble systems can be traced back to some earlier studies such as [12,13],it is Schapire's 1990 paper[14] that is widely recognized as the seminal work on ensemble systems. Few more references for data fusions and combining classifiers are available in [15-21]. In this paper, study of ensemble of classifiers is presented using investigation on different datasets. It is important to submit here that there is a quite little scope of comparison of proposed scheme with others available in literature. The reason is that in each ensemble of classifiers, the constituent classifiers are well established classifiers, viz. neural networks, fuzzy, knn etc. The performances of these individual classifiers have already been widely reported in literature in several applications. For a simple implementation of the proposed scheme, k-NN classifier has been used in this paper as the constituent classifier of the ensemble. Presumably, the k-nearest neighbor algorithm[28] is considered one of the simplest machine learning algorithms. It is further to add that the objective here is not to discuss the strength of k-nn but to investigate the performance of the ensemble, irrespective of its consitiuent classifiers. However a good survey on k-nn classifier can be found at [31].

The objective of this paper is to support the creation of an ensemble with one or more of these classifiers as constituent members and to show that under an ensemble, the classifier accuracy produced by such an ensemble using majority of voting criterion, is always higher than that obtained by using individual classifier. This is supported by investigation on six benchmark datasets.

The paper is organized as follows: Section II presents proposed ensemble scheme. Section III outlines summary of datasets used in the experiments. The details of experiments and results are discussed in Section IV. Section V addresses the strength and weakness of proposed technique by comparing it with few of the others reported. Conclusions and future research prospects are reflected in Section VI followed by references.

2. PROPOSED ENSEMBLE ALGORITHM

In this paper, simple bagging without replacement of samples, with majority of voting [11,22,23] is used for the investigation of proposed scheme. Steps of the algorithm used in the paper are given below.

Algorithm

Input: *D,* the given dataset consisting of *N* patterns. *F,* number of features in each pattern, each pattern being labeled with a class *c* and *C* is the total number of classes in *D*. *S,* is number of classifiers in the ensemble.

1. Partition the entire dataset D into two parts, training dataset, S_{tr} and testing dataset, S_{te} . Each part has same number of features. Each pattern in these two parts is labeled with one class out of *C* classes, thus
$$S_{tr} \bigcup S_{te} = D$$
2. Make equal partitions of S_{tr} such that all parts except the last, will have S_{tr}/S patterns. The last part will have (S_{tr}/S+ S_{tr}%S), where % is modulus operation on integers. The ensemble will thus have *S* number of classifiers, one for each part.
3. Invoke k-nearest neighbor classifier [32] with k =1.
4. Determine the classification accuracy, C_a of each part of the training data using k-nn, against the same test data set S_{te}. Find out the average C_a of all *S* classifiers. Determine the maximum C_a obtained in the *S* classifiers.
5. Shuffle dataset *D;* create new S_{tr} and S_{te}.
6. Iterate steps 2 to 5, *I* times. Find the C_a and maximum $C_a$in these *I* iterations. .
7. Take every pattern of S_{te} and pass it through all *S* classifiers using bagging [11,22] and majority of voting techniques to determine it's class. Repeat the process for *I* times, Calculate average and maximum C_a of the ensemble (EC_a) in these *I* trials.
8. Change the value of *k* (1,3,5)
9. Change the value of *S* (5,7,9).
10. Change the size of training data (80%,60% and 50%) and accordingly test data.

Order of algorithm: There has been a variety of work in analysis of k nearest neighbors[34,35]. In the simplest form as used here [1,33], for k-nn, the order of search is $O(kdt_rt_e)$ where *F* is number of features (dimensions) in each pattern, *k* is number of nearest neighbors, Euclidean distance is used as a metric of nearest hood between test point t_e and training pattern t_r,, *P* is the preprocessing due to shuffling and partitioning of training (and testing) datasets, talking majority decision in bag of *S* classifiers For complete algorithm proposed here, order of algorithm may be given as follows

$$O(kFt_rt_e + P)$$

The algorithm is iterated for *k* as 1,3,5; *S* as 5,7,9; and size of training dataset as 80%,60% and 50%.

Fig. 1 shows the proposed scheme. In this figure, as a typical example, five classifiers are placed in an ensemble. The parts of training data $S_1 \ldots S_5$ are used for creating five classifiers $C_1 \ldots C_5$., one classifier for one part respectively. The CA of ensemble is shown by C_e.

3. SUMMARY OF DATABASES

Table 1summarizes data sets used for the experiments. The data sets are well established and have been used in several investigations. The details of each data set can be viewed in UCI Machine Learning Repository [24]. There has been no preference to choose any particular data set for investigation in this paper.

Data Set	Total Patterns	Features	Classes	Patterns in Class1	Patterns in Class2	Patterns in Class3
Iris	150	4	3	50	50	50
Wine	178	13	3	59	71	48
Liver	345	6	2	145	200	-
Thyroid	215	5	3	150	35	30
WBC	683	9	2	444	239	-
Sonar	208	60	2	97	111	-

Table 1:Description of the Data Set Used.

4. EXPERIMENTS AND RESULTS

Proposed ensemble algorithm was run on an i5 machine using MATLAB software. The purpose of the investigation was to focus the strength of proposed ensemble scheme over individual classifiers. The results obtained for the six classical databases are shown in Tables 2(A-F) for Iris, Wine, Bupa Liver, Thyroid, WBC(Wisconsin Breast Cancer) and Sonar datasets respectively. In each of these tables, first column: T, (training data size) indicates the part (in percent) of the database which will be used for training only whereas the remaining part (100 – T) will be used for testing. Three sizes for training have been used in the paper viz. 80%, 60% and 50%, to reflect attitude of the proposed algorithm towards different parts of the data. The next column represents values of 'k', i.e. the k-th nearest neighbor from the testing data pattern. The measure of the distance is taken as Euclidean distance. Three values of 'k' (1,3 and 5), have been used for all these datasets. To apply bagging, each training dataset is divided into *S* number of sub sets. In the paper, *S* is set for three values: 5,7 and 9. In other words, number of classifiers in ensemble will be 5, 7 and 9 for each of the datasets. Thus for each dataset, a training part of the dataset (80/60/50 %), has *S* different subsets. For a typical training dataset with five folds or subsets

$$S_1 \bigcup S_2 \bigcup S_3 \bigcup S_4 \bigcup S_5 = S_{tr}$$

and

$$S_{tr} \bigcup S_{te} = D$$

where S_{tr} and S_{te} stand for training and testing Dataset respectively, and D is the entire dataset. As a typical case, first experiment is conducted with S=5, k=1 and training data size =80% of the total dataset. The testing data (20%) will remain as the unseen part of dataset. In this case, each of these five classifiers, $S_1 \ldots S_5$,is applied to its respective training data part, e.g. first classifier accuracy *CA* will be obtained using 1-NN between $S_1$and test dataset, second *CA* between S_2same test data set and so on. The mean (average)of these five C_a is computed. The C_a of ensemble of classifier is computed as follows. Take first test pattern from test database and find its class using first nearest neighbor (1-NN) with S_1 then find its class with S_2,S_3,S_4,S_5. The majority (mode) of values of classes so obtained in five tests will be the class accepted for the ensemble. Repeat the exercise for all the patterns in test dataset and calculate its percentage C_a. This will compute *E* C_a of the ensemble. Time in execution of the whole process is also recorded. The whole exercise is repeated for five times by shuffling randomly the dataset. Compute the mean classification accuracy *Mean* C_a,, from these five iterations. Also calculate the maximum value of C_a , *Max* C_a in these five iterations. Similarly compute mean and maximum ensemble accuracy *Mean* EC_a and *MaxEC*$_a$ in the five iterations. Compute the mean time spent on one iteration. The mean values are shown in Tables 2(A-F). The maximum values for classification accuracies in five iterations are shown within the brackets in the same tables. This is shown by the first row of the first main sub column of Table with *S*=5. Similar exercise is repeated for *S*=7 and 9. This completes row 1 of the table. The values of k are varied to 3 and 5. Then training data size is changed to 60% and 50% and exactly same procedure is adopted. Due to space limitations, the values in tables are rounded up to two decimal places. The tables 2(A-F) are enclosed in Annexure-1.

On observing these Tables 2(A), it is noted that for iris data set, for S=5, k=1, *meanCa* =94.7 is highest when individual classifiers are considered. In this case *meanE* C_a is 96.7. For S=7, k=1, *mean* C_a =90.5 is highest for individuals, whereas *mean* C_a =96.6. For S=9, *mean* C_a =93.3, *mean* C_a =100%, Thus it is noted that *meanECA* is in each case is higher than *meanCA*. In most cases, mean C_a is same as maximum value of C_a. Typically, for 50% training data, S=7, k=1, *mean* C_a =85.7 whereas *max* C_a is 91.6. Similarly *meanE* C_a is 88.5 and *maxE* C_a =92.6. There are few more cases where mean values of CA are smaller than maximum values of C_a .Similar trend is noted in all tables 2(A-F).

As another case, Table 2(E) can be quoted which presents results on breast cancer (wbc) data. This dataset has 683 patterns divided into 444 and 239 patterns for class 1 and class2 respectively. Dataset has 9 features(attributes). With nine (S=9) 1-NN classifiers, *mean* C_a 96.6 whereas *max* C_a =97.9. The mean *E* C_a =98.5 with maximum *ECA* as 99.3%, a better performance. Sonar dataset has 208 patterns divided into two classes having 97 and 111 patterns respectively. It has 60 features in each pattern. By observing Table 2(C), it is noted that *mean* C_a =61.3, with 60% training data and 1-NN, using nine classifiers (S=9), whereas mean*E* C_a under similar conditions is 71.1.

It is therefore observed from study of all these tables that the values of mean C_a are always less than mean *E* C_a. The maximum values of C_a in few cases are greater than the mean values of C_a in five iterations. The reason for running experiments for five times is just to ensure that the performance of the classifiers can be checked under all possible patterns combinations in training and test datasets. It is again apprehended that each ensemble can contain any set of similar or combination of classifiers such as neural networks, fuzzy, Bayesian, kNN etc. The contribution of the paper is more towards showing the importance of the ensemble with majority of voting than to highlight the strength of constituent classifiers which are undoubtedly proven in

literature. That is why the classification accuracies of constituent classifiers are compared with that of the ensemble and not with other constituent classifiers of the ensemble. As a typical example kNN is used in all cases.

5. DISCUSSIONS ON THE COMPARATIVE STUDY OF PROPOSED TECHNIQUE

The proposed technique has been used for different applications e.g. in [36] , researchers used ensemble classifier for fMRI data analysis. There are various strong merits of the proposed scheme including high possibility of getting better classification accuracy from an ensemble than an individual classifier; the individual classifiers of the ensemble need not to be perfectly trained, mostly these are weak learners, thereby reducing the time and efforts of training them; the fact is also confirmed when different sizes of the training dataset is taken (80%,60% and 50%) still a good accuracy is achieved; there is a scope for feature selection and dimensionality reduction of the dataset, under different combinations of features, the ensemble can be called to predict a reasonable good accuracy.

Although it is difficult to find a common platform to compare the performance of proposed technique with some other used in different context, yet few results are being discussed here for the purpose

For iris data, the accuracy obtained in [37] is 94.7 for CBA scheme 96.6 for Neural Network system, where as with the proposed technique it is 100% for 9 classifiers in the ensembles with 80% training data for validation with k as 1.

For thyroid data [7], the accuracy is 95% with time as 0.913 seconds. In proposed scheme, the accuracy is 95.4 with S=4, k=1, time = 0.50 seconds.

For wine dataset, accuracy in [7] is although 89% but time taken is 1.34 seconds whereas in proposed scheme accuracy is 81.7 but time is quite less, 0.53 seconds ($k=1,S=9$, training data T = 60%).

For WBC data, in [38], the classification accuracy is 90% with time taken is 48 seconds whereas in the proposed technique, the accuracy is 98% ($k=1,S=9$, training data T = 80%) with time = 1.6 seconds.

For sonar data, the accuracy obtained in [39], is 81% whereas the accuracy obtained by proposed technique is approximately 79% ($k=1,S=5$, training data T = 80%).

It is again reminded that the proposed technique focuses on the use and importance of an ensemble of classifiers and not of an individual classifier.

One possible inability of the proposed technique is that it does not address or attempt to modify the original structure of any individual constituent classifiers. If a classifier originally does not fit suitable for a particular dataset or on a specific nature of data, the ensemble by no means will be able to improve its performance. Moreover for a large set of data such as micro array gene data, the performance of the proposed technique is subject to test.

6. CONCLUSION

In this paper, a recent yet important scheme of classification has been presented. A classifier can produce good classification accuracy for one dataset, but performs poorer when presented with different dataset or even different section of the same dataset. If however, multiple classifiers are trained for small sections of the databases, and are combined in the form of an ensemble, then such an ensemble can produce better classification accuracy. To justify it, six bench mark datasets, iris, BUPA liver, thyroid, sonar, breast cancer and wine have been used for empirical study. The size of the training part of each dataset is taken as. 80%,60% and 50%. The number of classifiers in the ensemble is taken as 5,7 and 9. The k-nearest neighbor has been used as classifier with the values of k as 1,3 and 5 under each case. Experiments were conducted to evaluate the classification accuracies of all six datasets. In order to provide diversity in training and testing datasets, the experiments were iterated for five times with shuffling of dataset. The mean and the maximum classification accuracies of individual classifiers on each sub sets of training datasets were computed. The same were computed for ensemble of the classifiers using majority of voting. The results produced in these two cases, show that the classification accuracy of each individual classifier in general is lower than that of the classification accuracy obtained by their ensemble. Thus it is concluded from these investigations that an ensemble is a good approach to determine the class of an unseen data pattern. The scheme can be applied to many other datasets. Moreover, other classifiers like neural network, fuzzy etc. can be included in the ensemble. This study can also be extended to explore the possibility of feature selection or dimensionality reduction.

REFERENCES

[1]. Duda, R.O., Hart, P.E., Stork, D.G., *Pattern Classifi*cation. John Wiley and Sons (Asia) Pte.Ltd., second .ed. 2006.

[2]. Kamber, M., Han, J.,Pei,*Data mining: Concepts and techniques*,2nd ed. CA: Morgan Kaufmann Publisher. San Francisco, 2011.

[3]. Jean-Marc Adamo, Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms, Springer, 2001.

[4]. Misra, B.B., Dehuri, S., Dash, P.K., Panda, G., "Reduced Polynomial Neural Swarm Net for Classification Task in Data Mining", IEEE Congress on Evolutionary Computation,2008b

[5]. Kosala, R., Blockeel, H., "Mining Research: A Survey". ACM SIGKDD Explorations. 2 (1), 2000, pp 1–15,.

[6]. Baldi, P., Brunak, S., *Bioinformatics: The Machine Learning Approach.* MIT Press, Cambridge, MA.,1998.

[7]. Saxena, Patre,Dubey, "An Evolutionary Feature Selection Technique Using Polynomial Neural Network", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): pp. 1694-0814 www.IJCSI.org.

[8]. Mitchel, T.M.,*Machine Learning*. McGraw Hill,1997.

[9]. Polikar R, “Ensemble based systems in decision making”, IEEE Circuits Syst Mag 6(3):pp. 21–45, 2006.

[10]. Mikel Galaretal. , “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches”, IEEE Transaction on Systems, Man and Cybernetics, Part-C,Applications and reviews, pp. 1-22, 2011.

[11]. Polikar,“Bootstrap methods”, in IEEE Signal Processing Magazine,July 2007.

[12]. B.V. Dasarathy and B.V. Sheela, “Composite classifier system design: Concepts and methodology,” Proc. IEEE, vol. 67, no. 5, pp. 708–713, 1979.

[13]. L.K. Hansen and P. Salamon, “Neural network ensembles,” IEEE Trans. Pattern Anal. Machine Intell., vol. 12, no. 10, pp. 993–1001, 1990.

[14]. R.E. Schapire, “The strength of weak learnability,” Machine Learning, vol. 5,no. 2, pp. 197–227, June 1990.

[15]. Okun, Oleg,Supervised and Unsupervised Ensemble Methods and their Applications, Springer, 2008,

[16]. Oleg Okun, *Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations*. , IGI Global, Hershey, PA, 2011.

[17]. Verma, B.,”Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning”, IEEE Transactions on Knowledge and Data Engineering, vol 24(4), pp 605-618, 2011.

[18]. E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” Machine Learning, vol. 36, no. 1-2, pp. 105–139,1999.
[Online].Available:http://citeseer.ist.psu.edu/bauer99empirical.html

[19]. D. Ruta and B. Gabrys, “An overview of classifier fusion methods,” Computing and Information Systems, vol. 7, pp. 1–10, 2000.

[20]. A. Al-Ani and M. Deriche, “A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence,” Journal of Artificial Intelligence Research, vol. 17, pp. 333–361, 2002.

[21]. L. K. Hansen and P. Salomon, “Neural network ensembles,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12(10), pp. 993–1001, October 1990.

[22]. Lior Rokach, “Ensemble-based classifiers”, Artificial Intelligence Rev pp. 33:1–39, Springer, 2010.

[23]. Saxena,Mondol,Mir, “Improving the Classification ccuracy with Ensemble of Classifiers”, proc. Of National Conference NCMIRA 12, India,21-23 Dec 2012, pp. 52-56.

[24]. http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html

[25]. Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics 7 (2): 179–188.

[26]. Domingos, Pedro & Michael Pazzani (1997) "On the optimality of the simple Bayesian classifier under zero-one loss". Machine Learning, 29:103–137

[27]. Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20, issue 3, 273-297, 1995

[28]. Cover TM, Hart PE, "Nearest neighbor pattern classification". IEEE Transactions on Information Theory 13 (1): 21–27, 1967.

[29]. Haykin, S. , Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.

[30]. Angelov P., X. Zhou, “Evolving Fuzzy-Rule-based Classifiers from Data Streams”, IEEE Transactions on Fuzzy Systems, ISSN 1063-6706, special issue on Evolving Fuzzy Systems, December 2008, vol. 16, No6, pp.1462-1475.

[31]. Liangxiao Jiang, Wuhan Zhihua Cai; Dianhong Wang; Siwei Jiang, “Survey of Improving K-Nearest-Neighbor for Classification”, in Proc. of IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007. Vol.1, pp. 679-683.

[32]. www.mathworks.in/help/stats/knnsearch.html

[33]. Machine Learning, Tom Mitchell, McGraw Hill, 1997

[34]. Gayathri, K., Marimuthu, A.; “Text document pre-processing with the KNN for classification using the SVM” ,in Proc. Of 7th International IEEE Conference on Intelligent Systems and Control (ISCO), 2013, pp. 453-457.

[35]. Efendi Nasibov, , Cagin Kandemir-Cavas; “Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction”, Computational Biology and Chemistry, Elsevier, vol 33(6), 2009, pp. 461, 464.

[36]. Ludmila I. Kunchevaa,, Juan J. Rodríguezb, “Classifier ensembles for f MRI data analysis: an experiment”, Elsevier, Magnetic Resonance Imaging,28(2010), 583-593.

[37]. Prachitee Shekhawat1,Sheetal S. Dhande, “Building an Iris Plant Data Classifier Using Neural Network Associative Classification”, International Journal of Advancements in Technology http://ijict.org/, vol 2(4), 2011,pp 491-506.

[38]. Saxena,Patre,Dubey, “Investigating a novel GA-based feature selection method using improved KNN classifiers”, Int. J. Information and Communication Technology, Vol. 3, No. 3, 2011,pp 274-288.

[39]. Saxena,Pal,Kothari, “Evolutionary methods for unsupervised feature selection using Sammon’s stress function” , Fuzzy Information and Engineering Volume 2, Number 3, 229-247, DOI: 0.1007/s12543-010-0047-4, Springer, http://www.springerlink.com/content/1616-8658/2/3/

[40]. P. C. Jha, Shivani Bali and P. K. Kapur, “Fuzzy Approach for Selecting Optimal COTS Based Software Products Under Consensus Recovery Block Scheme”,

BIJIT - BVICAM's International Journal of Information Technology, Vol. 3(1), 2011

Annexure-1

Table 2: Results obtained for six datasets used in ensemble of classifiers

(A) Iris data

T	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a(MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	MeanTime	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	94.7(94.7)	96.7(96.7)	0.43	90.5(90.5)	96.6(96.6)	0.48	93.3(93.3)	100(100)	0.47
	3	88.7(88.7)	96.7(96.7)	0.46	89.0(89.0)	90.0(90.0)	0.49	92.5(92.5)	100(100)	0.48
	5	94.0(94.0)	96.7(96.7)	0.47	88.5(88.5)	93.3(93.3)	0.50	83.7(83.7)	96.6(96.6)	0.51
60	1	94.7(94.7)	98.3(98.3)	0.50	89.3(89.3)	95.0(95.0)	047	90.5(90.5)	96.7(96.7)	0.51
	3	91.4(91.4)	98.4(98.4)	0.49	85.0(85.0)	91.7(91.7)	052	79.5(79.5)	95.0(95.0)	0.52
	5	89.4(89.4)	91.7(91.7)	0.49	80.5(80.5)	95.0(95.0)	052	62.5(62.5)	93.4(93.4)	056
50	1	91.7(91.7)	98.7(98.7)	0.49	**85.7(91.6)**	**88.5(92.6)**	0.47	86.9(90.1)	94.4(94.6)	0.44
	3	86.1(86.1)	86.7(86.7)	0.46	80.2(80.6)	93.6(96.0)	0.52	77.9(82.8)	91.7(92.0)	0.53
	5	78.4(78.4)	96.0(96.0)	0.49	62.5(68.6)	81.3(86.6)	0.51	66.5(69.3)	81.9(86.6)	0.55

(B) liver data

T	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	MeanTime	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	59.1(63.8)	66.9(73.9)	0.73	56.6(56.7)	60.9(60.9)	0.69	53.9(56.8)	60.6(71.0)	0.81
	3	61.5(63.2)	66.9(69.6)	0.73	59.3(59.6)	61.7(62.3)	0.74	60.1(60.1)	68.2(68.2)	0.68
	5	59.8(61.2)	62.9(65.2)	0.76	59.1(59.1)	71.0(71.0)	0.74	60.5(63.6)	66.4(71.0)	0.72
60	1	55.5(57.1)	59.3(64.5)	0.73	54.7(57.1)	60.6(62.3)	0.71	56.7(56.7)	59.4(59.4)	0.68
	3	57.1(58.4)	60.7(63.1)	0.69	58.7(59.1)	65.5(65.9)	0.72	59.0(59.7)	67.9(68.8)	0.74
	5	58.9(60.0)	61.6(63.0)	0.69	60.1(60.7)	66.2(66.6)	0.72	58.4(58.4)	64.5(64.5)	0.69
50	1	56.4(56.9)	61.5(65.7)	0.83	55.3(59.3)	61.3(69.2)	0.73	57.5(57.7)	65.5(69.2)	0.73
	3	60.8(60.8)	67.4(67.4)	0.72	58.1(58.6)	65.2(66.9)	0.72	57.4(57.9)	69.1(72.1)	0.78
	5	56.5(60.8)	59.8(69.2)	079	57.9(59.4)	64.5(66.9)	0.73	56.9(57.6)	64.4(66.2)	0.86

(C) sonar data

T	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a(MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	70.3(71.4)	78.1(78.6)	0.79	66.1(68.4)	69.5(73.8)	0.78	66.9(66.9)	77.6(78.6)	0.72
	3	65.5(66.6)	72.4(74.9)	0.76	57.3(64.9)	58.6(73.8)	0.80	48.5(56.3)	47.1(64.3)	0.76
	5	60.9(60.9)	66.6(66.6)	0.76	59.9(60.5)	69.0(69.0)	0.76	54.4(56.3)	58.1(61.9)	0.77
60	1	65.2(69.4)	71.9(74.7)	0.73	61.3(61.9)	66.7(67.4)	0.76	61.3(61.3)	71.1(71.1)	0.80
	3	59.0(59.3)	62.2(65.1)	0.75	55.6(55.6)	60.2(60.2)	0.79	57.0(58.4)	63.1(63.9)	0.76
	5	57.3(60.2)	61.4(61.4)	0.74	55.4(55.9)	63.4(66.2)	0.74	54.2(55.2)	59.8(62.7)	0.78
50	1	65.6(65.6)	69.2(69.2)	0.73	62.7(63.8)	73.1(75.0)	0.77	59.9(60.7)	68.1(72.1)	0.75
	3	59.1(59.4)	61.5(64.4)	0.72	57.1(58.2)	63.8(64.4)	0.77	56.8(57.2)	67.5(69.2)	0.78
	5	58.7(59.6)	61.5(63.5)	0.75	51.4(51.4)	50.9(50.9)	0.74	56.4(56.7)	62.9(72.1)	0.80

(D) thyroid data

T	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a(MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	MeanTime	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	90.2(90.2)	95.4(95.4)	0.50	87.4(87.4)	93.0(93.0)	0.55	77.9(79.3)	86.5(88.4)	0.53
	3	86.5(86.5)	88.4(88.4)	0.53	83.0(83.0)	90.7(90.7)	0.59	79.6(81.9)	80.9(83.7)	0.62
	5	83.3(83.3)	86.1(86.1)	0.54	74.2(76.4)	72.5(76.7)	0.53	77.4(77.7)	69.8(69.8)	0.59
60	1	82.6(92.0)	84.4(96.5)	0.55	76.4(76.4)	81.4(81.4)	0.53	80.2(82.8)	79.7(80.2)	0.57
	3	79.5(79.5)	82.5(82.5)	0.51	75.6(75.9)	76.7(76.7)	0.56	71.6(72.0)	68.1(70.9)	0.64
	5	78.1(78.1)	80.2(80.2)	0.58	73.0(73.0)	72.1(72.1)	0.51	74.3(75.3)	70.0(70.9)	0.62
50	1	81.3(81.3)	81.3(81.3)	0.57	86.8(86.8)	92.5(92.5)	0.55	80.7(80.7)	85.4(85.9)	0.61
	3	81.1(81.1)	85.0(85.0)	0.55	79.0(79.8)	79.8(80.4)	0.59	72.0(74.1)	67.5(71.9)	0.62
	5	78.5(78.5)	82.2(82.2)	0.57	70.4(70.7)	66.9(67.3)	0.59	80.4(80.4)	77.6(77.6)	0.59

(E) wbc data

T %	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a(MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	MeanTime	MeanE C_a (MaxE C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	95.4(97.2)	96.5(98.5)	1.35	95.5(97.8)	96.2(98.5)	1.50	96.6(97.9)	98.5(99.3)	1.60
	3	96.9(97.4)	96.6(97.0)	1.59	96.9(97.5)	97.5(98.5)	1.71	94.4(94.9)	95.2(96.3)	1.48
	5	95.7(96.5)	96.3(97.0)	1.53	96.5(97.7)	96.6(97.8)	1.43	96.3(97.2)	96.6(97.8)	1.51
60	1	95.3(96.3)	96.5(97.4)	1.55	94.9(96.0)	96.2(96.7)	1.66	95.8(96.6)	98.3(99.2)	1.54
	3	96.5(97.6)	97.4(98.5)	1.55	95.6(96.0)	96.5(97.4)	1.95	94.9(95.6)	95.0(98.2)	1.51
	5	95.7(98.1)	95.9(98.5)	1.56	96.5(97.2)	96.3(97.0)	1.66	94.5(96.0)	94.9(95.9)	1.71
50	1	94.6(94.9)	95.3(95.3)	1.42	95.1(95.4)	95.6(96.2)	1.63	94.3(95.0)	95.8(96.7)	1.71
	3	96.1(96.9)	96.5(97.4)	1.78	95.5(96.8)	96.2(97.4)	1.62	94.4(95.5)	94.9(95.6)	1.92
	5	95.6(96.3)	95.9(96.2)	1.72	95.6(95.8)	95.8(95.9)	1.50	96.0(96.6)	96.0(96.5)	1.63

(F) Wine data

T %	k	S=5			S=7			S=9		
		Mean C_a (Max C_a)	MeanE C_a(MaxE C_a)	Mean Time	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	MeanTime	Mean C_a (Max C_a)	MeanE C_a (MaxE C_a)	Mean Time
80	1	60.4(66.6)	58.3(69.4)	0.50	67.2(69.4)	76.6(83.3)	0.49	62.9(62.9)	75.0(75.0)	0.56
	3	64.6(65.0)	65.0(66.6)	0.54	64.5(68.5)	68.3(75.0)	0.57	70.1(70.1)	77.7(77.7)	0.61
	5	65.7(67.7)	68.8(72.2)	0.52	65.4(66.3)	60.5(61.1)	0.55	70.0(70.0)	75.0(75.0)	0.61
60	1	63.5(67.6)	67.6(73.2)	0.53	68.8(68.8)	77.5(78.8)	0.51	70.7(70.7)	81.7(81.7)	0.53
	3	67.0(67.3)	66.5(67.6)	0.55	64.5(69.2)	69.3(76.1)	0.60	61.6(61.6)	64.7(64.7)	0.57
	5	69.8(70.7)	73.8(76.0)	0.54	70.9(72.6)	74.3(76.0)	0.60	57.1(57.1)	61.9(61.9)	0.60
50	1	61.9(63.8)	64.7(67.4)	0.50	65.5(65.6)	69.6(69.6)	0.57	61.2(61.2)	67.4(67.4)	0.55
	3	69.2(70.5)	73.4(74.1)	0.55	59.7(66.3)	63.3(69.6)	0.57	62.5(62.5)	69.6(69.6)	0.62
	5	65.1(71.4)	74.6(75.2)	0.55	66.1(66.1)	70.7(70.7)	0.62	59.2(59.8)	65.1(66.2)	0.61

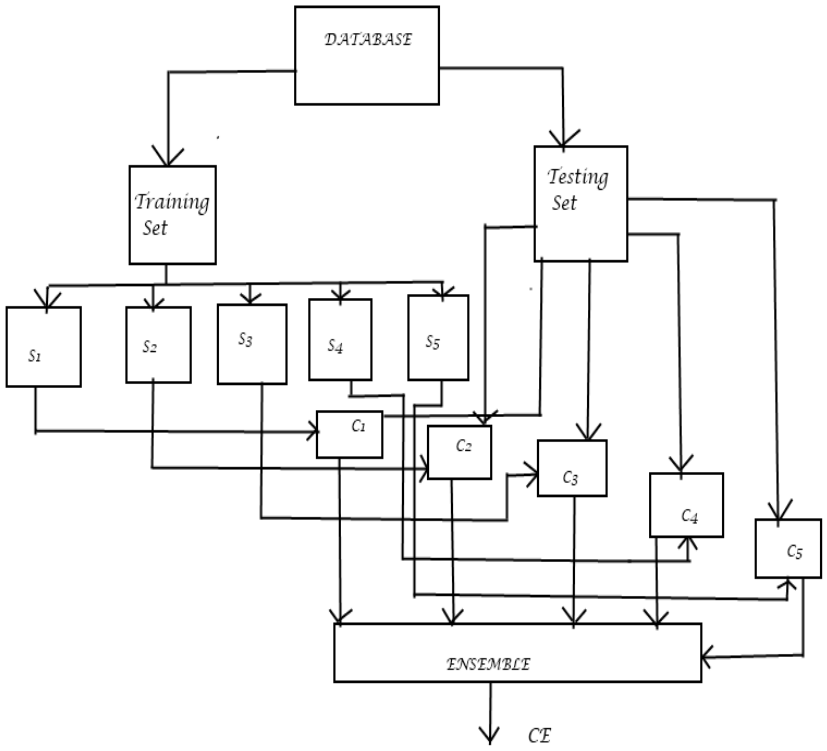


Figure1: "Representation of ensemble algorithm for number of classifiers, S=5"

BIJIT - BVICAM's International Journal of Information Technology

ISSN 0973 – 5658; Visit us at www.bvicam.ac.in/bijit/

Call for Papers

Special Issue on Fuzzy Logic

Submission Deadline: 31st August, 2013

Editor-in-Chief
Prof. M. N. Hoda
Director, BVICAM, New Delhi (INDIA)
E-Mail: mca@bvicam.ac.in

Guest Editor
Prof. M. M. Sufyan Beg
Professor, Dept. of Computer Engineering
Jamia Millia Islamia, New Delhi (INDIA)
E-Mail: mmsbeg@hotmail.com

BIJIT is a peer reviewed refereed bi-annual research journal having ISSN 0973-5658, being published since 2009, in both, Hard Copy as well as Soft copy. Two issues; **January – June** and **July – December**, are published every year. The journal intends to disseminate original scientific research and knowledge in the field of, primarily, Computer Science and Information Technology and, generally, all interdisciplinary streams of Engineering Sciences. In a short span of four years, BIJIT has been indexed with some of the world's leading indexing / bibliographic agencies like **EBSCO** (USA), **Open J-Gate** (USA), **DOAJ** (Sweden), **Google Scholar, WorldCat** (USA), **Cabell's Directory** of Computer Science and Business Information System (USA), **Academic Journals Database**, **Open Science Directory, Indian Citation Index, etc.** and listed in the libraries of the world's leading Universities like **Stanford University, Florida Institute of Technology, University of South Australia, University of Zurich**, etc. Related links are available at http://www.bvicam.ac.in/bijit/indexing.asp

BIJIT Seeks original and unpublished research papers for a Special Issue of the Journal on Fuzzy Logic scheduled to appear in March, 2014 Vol. 6, No. 1, January – June, 2014.

Over the years, Fuzzy Logic has changed our lives in a significant way. In a narrow-sense, Fuzzy Logic is considered to be a logical system, which is a generalization of multi-valued logic. A very important distinguishing feature of Fuzzy logic is that in Fuzzy logic, everything is, or is allowed to be, a matter of degree. Furthermore, the degrees are allowed to be fuzzy. In a broader sense, however, Fuzzy Logic is much more than a logical system. In fact, Fuzzy Logic is a precise system of reasoning and computation in which the objects of reasoning and computation are classes with unsharp boundaries. What is not widely recognized, within the scientific community and the general public, is that Fuzzy Logic has become a vast enterprise. There are over 280,000 papers in the literature with Fuzzy in title. There are 25 journals with fuzzy in title. There are close to 25,000 Fuzzy-Logic -related patents issued or applied for in the United States and Japan. There is a long list of applications ranging from digital cameras to fraud detection systems. Particularly worthy of note, on one end, is the Fuzzy Logic subway system in Sendai, a city of over 1 million in Japan. On the other end, numerically, is Omron's 120 million fuzzy logic blood pressure meters.

Some critics have been saying that Fuzzy Logic is a passing fad. This assessment of Fuzzy Logic fails to recognize that the world we live in is, in large measure, a world of Fuzzy classes, and that science has much to gain from shifting its foundation from classical Aristotelian logic to fuzzy logic.

It is on the above note that we wish to bring out a special issue, to celebrate the **Golden Jubilee Year of Fuzzy Logic,** in the year 2014 – the 50th year of the introduction of Fuzzy Logic by its Father, **Lotfi A. Zadeh,** in the year 1965.

Original and unpublished research papers, based on theoretical or experimental works, are solicited for publication in the Special Issue of BIJIT. Submission of a paper implies that the work described has not been published previously (except in the form of an abstract or academic thesis) and is not under consideration for publication elsewhere. Papers can be submitted electronically, after logging in at our portal and accessing the submit paper link, available at http://www.bvicam.ac.in/bijit/SubmitPaper.asp upto **31st August, 2013**, with **"Special Issue of BIJIT on Fuzzy Logic"** being selected as **Publication Type**. *E-Mailic submission will not serve the purpose.* Authors wishing to submit the paper to this Special Issue must refer to the website, for paper structuring and formatting guidelines in detail, at http://www.bvicam.ac.in/bijit/Basic_Guidelines_for_Authors.asp.

BIJIT follows double blind peer review system. All submitted papers are first assessed at editorial board level on the basis of their technical suitability, scope of work and plagiarism. The corresponding authors of qualifying submissions will be intimated for their papers to be double blind reviewed by at-least two experts on the basis of originality, novelty, clarity, completeness, relevance, significance and research contribution. If recommended, the paper may undergo multiple cycles of review, before finally being accepted. Final acceptance is based on the review remarks by the referees and decision of the editorial board. Publication of papers in BIJIT is **FREE OF COST**. We do not charge any **publication fee** from the authors for the papers to be published in BIJIT.

Timeline for Special Issue

Submission Deadline	**:**	**31st August, 2013**
First Notification	**:**	**31st October, 2013**
Author Revision Due	**:**	**18th November, 2013**
Notification of Acceptance, if Major Revision Required	**:**	**31st December, 2013**
Accepted Papers Due for Editorial Review	**:**	**10th January, 2014**
Final Acceptance Notification	**:**	**31st January, 2014**
Tentative Date of Publication	**:**	**March, 2014**

BIJIT - BVICAM's International Journal of Information Technology

Paper Structure and Formatting Guidelines for Authors

BIJIT is a peer reviewed refereed bi-annual research journal having **ISSN 0973-5658**, being published since 2009, in both, Hard Copy as well as Soft copy. Two issues; **January – June** and **July – December**, are published every year. The journal intends to disseminate original scientific research and knowledge in the field of, primarily, Computer Science and Information Technology and, generally, all interdisciplinary streams of Engineering Sciences. **Original** and **unpublished** research papers, based on theoretical or experimental works, are published in BIJIT. We publish two types of issues; **Regular Issues** and **Theme Based Special Issues.** Announcement regarding special issues is made from time to time, and once an issue is announced to be a Theme Based Special Issue, Regular Issue for that period will not be published.

Papers for Regular Issues of BIJIT can be submitted, round the year. After the detailed review process, when a paper is finally accepted, the decision regarding the issue in which the paper will be published, will be taken by the Editorial Board; and the author will be intimated accordingly. However, for Theme Based Special Issues, **time bound Special Call for Papers** will be announced and the same will be applicable for that specific issue only.

Submission of a paper implies that the work described has not been published previously (except in the form of an abstract or academic thesis) and is not under consideration for publication elsewhere. The submission should be approved by all the authors of the paper. If a paper is finally accepted, the authorities, where the work had been carried out, shall be responsible for not publishing the work elsewhere in the same form. *Paper, once submitted for consideration in BIJIT, cannot be withdrawn unless the same is finally rejected*.

1. Paper Submission

Authors will be required to submit, MS-Word compatible (.doc, .docx), papers electronically *after logging in at our portal and accessing the submit paper link*, available at http://www.bvicam.ac.in/bijit/SubmitPaper.asp. Once the paper is uploaded successfully, our automated Paper Submission System assigns a Unique Paper ID, acknowledges it on the screen and also sends an acknowledgement email to the author at her / his registered email ID. Consequent upon this, the authors can check the status of their papers at the portal itself, in the Member Area, after login, and can also submit revised paper, based on the review remarks, from member area itself. The authors must quote / refer the paper ID in all future correspondences. Kindly note that we do not accept E-Mailic submission. To understand the detailed step by step procedure for submitting a paper, click at http://www.bvicam.ac.in/BIJIT/guidelines.asp.

2. Paper Structure and Format

While preparing and formatting papers, authors must confirm to the under-mentioned MS-Word (.doc, .docx) format:-

- The total length of the paper, including references and appendices, must not exceed **six (06) Letter Size pages**. It should be typed on one-side with double column, single-line spacing, 10 font size, Times New Roman, in MS Word.
- The Top Margin should be 1", Bottom 1", Left 0.6", and Right 0.6". Page layout should be portrait with 0.5 Header and Footer margins. Select the option for different Headers and Footers for Odd and Even pages and different for First page in Layout (under Page Setup menu option of MS Word). Authors are not supposed to write anything in the footer.
- The title should appear in single column on the first page in 14 Font size, below which the name of the author(s), in bold, should be provided centrally aligned in 12 font size. The affiliations of all the authors and their E-mail IDs should be provided in the footer section of the first column, as shown in the template.
- To avoid unnecessary errors, the authors are strongly advised to use the "spell-check" and "grammar-check" functions of the word processor.
- The complete template has been prepared, which can be used for paper structuring and formatting, and is available at http://www.bvicam.ac.in/BIJIT/Downloads/Template_For_Full_Paper_BIJIT.pdf.
- The structure of the paper should be based on the following details:-

Essential Title Page Information

- **Title:** Title should be Concise and informative. Avoid abbreviations and formulae to the extent possible.
- **Authors' Names** and **Affiliations:** Present the authors' affiliation addresses (where the actual work was done) in the footer section of the first column. Indicate all affiliations with a lower-case superscript letter immediately after the author's name

and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and e-mail address of each author.

- **Corresponding Author:** Clearly indicate who will handle correspondence at all stages of refereeing and publication. Ensure that phone numbers (with country and area code) are provided, in addition to the e-mail address and the complete postal address.

Abstract

A concise abstract not exceeding 200 words is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. References and non-standard or uncommon abbreviations should be avoided. As a last paragraph of the abstract, 05 to 10 Index Terms, in alphabetic order, under the heading Index Terms ***(Index Terms - …….)*** must be provided.

NOMENCLATURE

Define all the abbreviations that are used in the paper and present a list of abbreviations with their definition in Nomenclature section. Ensure consistency of abbreviations throughout the article. Do not use any abbreviation in the paper, which has not been defined and listed in Nomenclature section.

Subdivision - numbered sections

Divide paper into numbered Sections as 1, 2, 3, ….. and its heading should be written in CAPITAL LETTERS, bold faced. The subsections should be numbered as 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. and its heading should be written in Title Case, bold faced and should appear in separate line. The Abstract, Nomenclature, Appendix, Acknowledgement and References will not be included in section numbering. In fact, section numbering will start from Introduction and will continue till Conclusion. All headings of sections and subsections should be left aligned.

INTRODUCTION

State the objectives of the work and provide an adequate background, with a detailed literature survey or a summary of the results.

Theory/Calculation

A Theory Section should extend, not repeat the information discussed in Introduction. In contrast, a Calculation Section represents a practical development from a theoretical basis.

RESULT

Results should be clear and concise.

DISCUSSION

This section should explore the importance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate.

CONCLUSION AND FUTURE SCOPE

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

APPENDIX

If there are multiple appendices, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similar nomenclature should be followed for tables and figures: Table A.1; Fig. A.1, etc.

ACKNOWLEDGEMENT

If desired, authors may provide acknowledgements at the end of the article, before the references. The organizations / individuals who provided help during the research (e.g. providing language help, writing assistance, proof reading the article, sponsoring the research, etc.) may be acknowledged here.

REFERENCES

Citation in text

Please ensure that every reference cited in the text is also present in the reference list (and vice versa). The references in the reference list should follow the standard IEEE reference style of the journal and citation of a reference.

Web references

As a minimum, the full URL should be given and the date when the reference was last accessed. Any further information, if known (DOI, author names, dates, reference to a source publication, etc.), should also be given. Web references can be listed separately (e.g., after the reference list) under a different heading if desired, or can be included in the reference list, as well.

Reference style

Text: Indicate references by number(s) in square brackets in line with the text. The actual authors can be referred to, but the reference number(s) must always be given. Example: '..... as demonstrated [3,6]. Barnaby and Jones [8] obtained a different result ....'

List: Number the references (numbers in square brackets) in the list, according to the order in which they appear in the text.

Two sample examples, for writing reference list, are given hereunder:-

Reference to a journal publication:

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread-spectrum watermarking for multimedia", *IEEE Transactions on Image Processing*, Vol. 6, No. 12, pp. 64 – 69, December 1997.

Reference to a book:

[2] J. G. Proakis and D. G. Manolakis – Digital Signal Processing – Principles, Algorithms and Applications; Third Edition; Prentice Hall of India, 2003.

Mathematical Formulae

Present formulae using Equation editor in the line of normal text. Number consecutively any equations that have to be referred in the text

Captions and Numbering for Figure and Tables

Ensure that each figure / table has been numbered and captioned. Supply captions separately, *not attached to the figure*. A caption should comprise a brief title and a description of the illustration. Figures and tables should be numbered separately, but consecutively in accordance with their appearance in the text.

3. Style for Illustrations

All line drawings, images, photos, figures, etc. will be published in black and white, in Hard Copy of BIJIT. Authors will need to ensure that the letters, lines, etc. will remain legible, even after reducing the line drawings, images, photos, figures, etc. to a two-column width, as much as 4:1 from the original. However, in Soft Copy of the journal, line drawings, images, photos, figures, etc. may be published in colour, if requested. For this, authors will need to submit two types of Camera Ready Copy (CRC), after final acceptance of their paper, one for Hard Copy (compatible to black and white printing) and another for Soft Copy (compatible to colour printing).

4. **Referees**

Please submit, with the paper, the names, addresses, contact numbers and e-mail addresses of three potential referees. Note that the editor has sole right to decide whether or not the suggested reviewers are to be used.

5. Copy Right

Copyright of all accepted papers will belong to BIJIT and the author(s) must affirm that accepted Papers for publication in **BIJIT** must not be re-published elsewhere without the written consent of the editor. To comply with this policy, authors will be required to submit a signed copy of Copyright Transfer Form, available at http://bvicam.ac.in/bijit/Downloads/BIJIT-Copyright-Agreement.pdf, after acceptance of their paper, before the same is published.

6. Final Proof of the Paper

One set of page proofs (as PDF files) will be sent by e-mail to the corresponding author or a link will be provided in the e-mail so that the authors can download the files themselves. These PDF proofs can be annotated; for this you need to download Adobe Reader version 7 (or higher) available free from http://get.adobe.com/reader. If authors do not wish to use the PDF annotations function, they may list the corrections and return them to BIJIT in an e-mail. Please list corrections quoting line number. If, for any reason, this is not possible, then mark the corrections and any other comments on a printout of the proof and then scan the pages having corrections and e-mail them back, within 05 days. Please use this proof only for checking the typesetting, editing, completeness and correctness of the text, tables and figures. Significant changes to the paper that has been accepted for publication will not be considered at this stage without prior permission. It is important to ensure that all corrections are sent back to us in one communication: please check carefully before replying, as inclusion of any subsequent corrections cannot be guaranteed. Proofreading is solely authors' responsibility. Note that BIJIT will proceed with the publication of paper, if no response is received within 05 days.

BIJIT - BVICAM's International Journal of Information Technology

(A Half Yearly Publication; ISSN 0973 - 5658)

Subscription Rates (Revised w.e.f. January, 2012)

	1 Year		3 Years	
Category	**India**	**Abroad**	**India**	**Abroad**
Companies	Rs. 1000	US $ 45	Rs. 2500	US $ 120
Institution	Rs. 800	US $ 40	Rs. 1600	US $ 100
Individuals	Rs. 600	US $ 30	Rs. 1200	US $ 075
Students	Rs. 250	US $ 25	Rs. 750	US $ 050
Single Copy	Rs. 500	US $ 25	-	-

Subscription Order Form

Please find attached herewith Demand Draft No._____ dated _____

For Rs._____ drawn on _____Bank in favor of **Director, "Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi"** for a period of 01 Year / 03 Years

Subscription Details

Name and Designation _____

Organization _____

Mailing Address _____

_____ PIN/ZIP _____

Phone (with STD/ISD Code)_____FAX_____

E-Mail (in Capital Letters)_____

Date: **Signature**

Place: (with official seal)

Filled in Subscription Order Form along with the required Demand Draft should be sent to the following address:-

Prof. M. N. Hoda
Editor-in- Chief, BIJIT
Director, Bharati Vidyapeeth's
Institute of Computer Applications & Management (BVICAM)
A-4, Paschim Vihar, Rohtak Road, New Delhi-110063 (INDIA).
Tel.: 91 – 11 – 25275055 Fax: 91 – 11 – 25255056 E-Mail: bijit@bvicam.ac.in
Visit us at: www.bvicam.ac.in/bijit

Announcement and Call for Papers

INDIACom-2014

8th INDIACom; 2014 International Conference on

Computing for Sustainable Global Development

(05th-07th March, 2014)

IEEE Conference Record Number # 32558

Bharati Vidyapeeth's
Institute of Computer Applications & Management (BVICAM)
A-4, Paschim Vihar, Rohtak Road, New Delhi-63 (INDIA)

Technically Sponsored by

IEEE Delhi Section

Supported by

The Institution of Electronics and Telecommunication Engineers (IETE), Delhi Centre

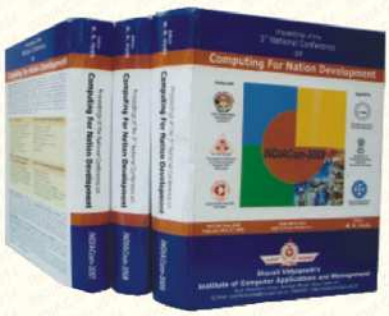
The Institution of Engineering and Technology Delhi Local Networks

CSI Region-I & CSI Division-I

GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY

IE (I), Delhi Centre

ISTE, Delhi Section



(Copies of the proceedings of past *INDIAComs*)

Correspondence

All correspondences related to the conference must be sent to the address:-

Prof. M. N. Hoda
General Chair, *INDIACom - 2014*
Director, BVICAM, A-4, Paschim Vihar, New Delhi -63 (INDIA)
Tel.: 91-11-25275055, TeleFax:91-11-25255056, 09212022066 (Mobile)
E-Mails: conference@bvicam.ac.in, indiacom2014@gmail.com
visit us at: http://www.bvicam.ac.in/indiacom

INDIACom-2014 is aimed to invite original research papers in the field of, primarily, Computer Science and Information Technology and, generally, all interdisciplinary streams of Engineering Sciences, having central focus on sustainable computing applications, which may be of some use in enhancing the quality of life and contribute effectively to realize the nations' vision of sustainable inclusive development using Computing. *INDIACom-2014* is an amalgamation of four different international conferences which will be organized parallel to each other, as parallel tracks. These are listed below:-

Track #1: International Conference on Sustainable Computing (ICSC-2014)
Track #2: International Conference on High Performance Computing (ICHPC-2014)
Track #3: International Conference on High Speed Networking & Information Security (ICHNIS-2014)
Track #4: International Conference on Software Engineering & Emerging Technologies (ICSEET-2014)

INDIACom-2014 will be held at **Bharati Vidyapeeth, New Delhi (INDIA)**. The conference will provide a platform for technical exchanges within the research community and will encompass regular paper presentation sessions, invited talks, key note addresses, panel discussions and poster exhibitions. In addition, the participants will be treated to a series of cultural activities, receptions and networking to establish new connections and foster everlasting friendship among fellow counterparts.

Full length original and unpublished research papers based on theoretical or experimental contributions related to the following topics, but not limited to, are solicited for presentation and publication in the conference:-

- Algorithms and Computational Mathematics
- Green Technologies and Energy Efficient Systems
- IT for Education, Health & Development
- IT for Environmental Sustainability
- IT for Sustainable Agriculture Development
- IT for Water Resources Management
- IT for Consumers' Right
- IT for Crisis Prevention & Recovery
- IT for Disaster Management and Remote Sensing
- IT for other day to day problems
- E-Governance
- Knowledge Management
- E-Commerce, ERP, CRM & Knowledge Mining
- Technology for Convergence
- Distributed and Cloud Computing
- Parallel, Multi-core and Grid Computing
- Reconfigurable Architectures
- Changing Software Architectural Paradigms
- Programming Practices & Coding Standards
- Software Inspection, Verification & Validation
- Software Sizing and Estimation Techniques
- Agile Technologies
- Artificial Intelligence and Neural Networks
- Computer Vision, Graphics, and Image Processing
- Modelling and Simulation
- Embedded Systems and Robotics
- Human Computer Interaction
- Databases
- Data Mining and Business Intelligence
- Big Data Analytics
- Operating Systems
- Data Communication, Computer Networks and Information Security
- Wireless Networking
- Network Monitoring Tools
- Next Generation Internet
- Mobile Computing
- Entertainment Technologies
- Multimedia Computing
- Information and Collaboration Systems
- Fuzzy and Soft Computing
- Bioinformatics
- Medical Informatics
- Education Informatics
- Computational Finance
- Research Methods for Computing
- Case Studies & Applications

Paper Submission

Authors from across different parts of the world are invited to submit their papers. Authors wishing to submit their papers must refer to the website, for paper structuring and formatting guidelines in detail, at http://www.bvicam.ac.in/indiacom/Technical%20Guidelines.asp. Authors should submit their papers online at http://www.bvicam.ac.in/indiacom/loginReqSubmitPaper.asp. Unregistered authors should first create an account on http://www.bvicam.ac.in/indiacom/addMember.asp to log on and submit paper. Only electronic submissions will be considered. E-Mailic submissions will not be considered.

Review Process, Publication and Indexing

The conference aims at carrying out two rounds of review process. In the first round, the papers submitted by the authors will be assessed on the basis of their technical suitability, scope of work and plagiarism. The corresponding authors of qualifying submissions will be intimated for their papers to be double blind reviewed by at-least two experts on the basis of originality, novelty, clarity, completeness, relevance, significance and research contribution. The shortlisted papers will be accepted for presentation and publication in the conference proceedings, having **ISSN 0973–7529** and **ISBN 978-93-80544-10-6** serials. Conference proceedings will also be available in the form of CD-ROMs. All accepted papers, which will be presented in the conference, will be submitted for inclusion to **IEEE Xplore,** as a part of **IEEE's Conference Publication Programme**, subject to their terms and conditions. Further details are available at www.bvicam.ac.in/indiacom.

Important Dates

Submission of Full Length Paper	04th November, 2013	**Paper Acceptance Notification**	13th January, 2014
Submission of Camera Ready Copy (CRC) of the Paper	20th January, 2014	**Registration Deadline** (for inclusion of Paper in Proceedings)	31st January, 2014