July – December, 2015 Vol. 7 No. 2 ISSN 0973-5658

BIJIT

Impact Factor 0.605 and Index Copernicus Value (ICV) 4.75

BVICAM's International Journal of Information Technology

Indexed with INSPEC (UK), Index Copernicus (Poland), ProQuest (UK), EBSCO (USA), Google Scholar (USA)

CONTENTS

BHARATI VIDYAPEETH PUNE

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

Volume 7, Number 2, July – December, 2015

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

Disclaimer
The opinions expressed and figures provided in the Journal; BIJIT, are the sole responsibility of the authors. The publisher and the editors bear no responsibility in this regard. Any and all such liabilities are disclaimed

All disputes are subject to Delhi jurisdiction only.

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 – 110063 (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

Our Major Indexing at International Level

The **INSPEC, IET (UK),** formerly IEE (UK), database is an invaluable information resource for all scientists and engineers, that contains 13 million abstracts and specialized indexing to the world's quality research literature in the fields of physics and engineering. *For further details, click at http://www.theiet.org/resources/inspec/*

Index Copernicus International (Poland) is a journal indexing, ranking and abstracting site. This service helps a journal to grow from a local level to a global one as well as providing complete web-based solution for small editorial teams. ICV 2012 for the BIJIT is 4.75. *For further details, click at http://jml2012.indexcopernicus.com/BVICAMs+International+Journal+of+Information+Technology,p4852,3.html*

ProQuest (UK) connects people with vetted, reliable information. Key to serious research, the company has forged a 70-year reputation as a gateway to the world's knowledge – from dissertations to governmental and cultural archives to news, in all its forms. *For further details, click at http://www.proquest.co.uk/en-UK/default.shtml*

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=*

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*

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*

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

Editorial Board

Prof. Kishor Trivedi
Dept. of Electrical and Computer Engineering , Duke University (USA)

Prof. Kukjin Chun
Dept. of Electrical and Computer Engineering, Seoul National University (KOREA)

Prof. M. N. Doja
Dept. of Computer Engineering, Jamia Millia Islamia, New Delhi (INDIA)

Prof. M. P. Gupta
Dept. of Management Studies, IIT Delhi, New Delhi (INDIA)

Prof. Madan Gupta
Director, Intelligent Systems Research Laboratory, University of Saskatchewan, Saskatoon, Saskatchewan (CANADA)

Dr. Nathalie Mitton
INRIA (FRANCE)

Dr. Nurul Fadly Bin Habidin
Engineering Business and Management, University Pendidikan Sultan Idris (MALAYSIA)

Prof. O. P. Vyas
Dept. of Information Technology, Indian Institute of Information Technology Allahabad (IIITA), Allahabad (INDIA)

Dr. Prabhaker Mateti
Dept. of Computer Science and Engineering, Wright State University (USA)

Prof. Prasant Mohapatra
Dept. of Computer Science, University of California (USA)

Prof. Richard Chbeir
School of Computer Science, Université de Pau et des Pays de l'Adour (UPPA), Anglet (FRANCE)

Dr. S. Arockiasamy
Dept. of Information Systems, University of Nizwa, Sultanate of Oman (OMAN)

Prof. S. I. Ahson
Former Pro-Vice-Chancellor, Patna University, Patna (INDIA)

Prof. S. K. Gupta
Dept. of Computer Science and Engineering, IIT Delhi, New Delhi (INDIA)

Prof. Salim Beg
Dept. of Electronics Engineering, Aligarh Muslim University, Aligarh (INDIA)

Prof. Shiban K. Koul
Centre for Applied Research in Electronics (CARE), IIT Delhi, New Delhi (INDIA)

Prof. Shuja Ahmad Abbasi
Dept. of Electrical Engineering, King Saud University, Riyadh (KSA)

Prof. Steven Guan
Dept. of Computer Science & Software Engineering, Xi'an Jiaotong-Liverpool University (CHINA)

Prof. Subir Kumar Saha
Dept. of Mechanical Engineering, IIT Delhi, New Delhi (INDIA)

Prof. Subramaniam Ganesan
Dept. of Computer Science and Engineering, Oakland University, Rochester (USA)

Prof. Susantha Herath
School of Electrical and Computer Engineering, St. Cloud State University, Minnesota (USA)

Prof. Yogesh Singh
Director, NSIT, New Delhi & Former Vice Chancellor, MS University, Baroda (INDIA)

Editorial

It is a matter of both honor and pleasure for us to put forth the fourteenth 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 again 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 ninety two 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 ***INSPEC*** *of IET (UK) formerly IEE (UK),* ***Index Copernicus International*** *(Poland) with* ***IC Value 4.75***, ***ProQuest*** *(UK),* ***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. Based upon the papers published in the year 2012, its* ***Impact Factor*** *was found to be* ***0.605****. These encouraging results 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

GA Based Clustering of Mixed Data Type of Attributes (Numeric, Categorical, Ordinal, Binary and Ratio-Scaled)

Rohit Rastogi[1], Pinki Mondal[2], Kritika Agarwal[3], Rachit Gupta[4] and **Shilpi Jain[5]**

Submitted in July 2013; Accepted in March, 2015

Abstract - Data mining discloses hidden, previously unknown, and potentially useful information from large amounts of data. As comparison to the traditional statistical and machine learning data analysis techniques, data mining emphasizes to provide a convenient and complete environment for the data analysis. Data mining has become a popular technology in analyzing complex data. Clustering is one of the data mining core techniques.

Data mining and data clustering, the prominent field of today it is a highly desirable task to apply unsupervised classification analysis on high volume of data sets with combined ordinal, ratio-scaled, binary and nominal with numeric, categorical, with values. However, most already available data merging and grouping through unsupervised classification algorithms are effective for the data with numeric category rather than the mixed data set. So,in this paper we have made efforts to present a new amalgamation techniques for these combined data sets by doing changes in the common cost function, and here we have tofindtraceof the internal cluster dispersion matrix.

To obtain correct clustering result the algorithm used is GA that optimizes the new cost function. We can compare and analyze that for high dimensional sets of data having mixed attributes GA-based clustering algorithm is feasible.

Core Idea of Our Paper

By this paper, we try to describe a technique for estimating the cost function metrics from mixed numeric, categorical and other type databases by using a uncertain grade-of-membership clustering model with the efficiency of Genetic Algorithm. This technique can be applied to the problem of opportunity analysis for business decision-making.

This general approach could be adapted to many other applications where a decision agent needs to assess the value of items from a set of opportunities with respect to a reference set representing its business. For processing numeric attributes, instead of generalizing them, a prototype may be developed for experiments with synthetic and real data sets,

and comparison with those of the traditional approaches. The results confirmed the feasibility of the framework and the superiority of the extended techniques.

[1]Sr. Asst Professor, CSE-Dept-ABES Engg. College, Ghaziabad (U.P.), India, +91-9818992772

[2,3,5]B.Tech.CSE-IIIYr., CSE-Dept.-ABES Engineering College, Ghaziabad (U.P.), India

[4]B.Tech. IT-Final Yr., IT-Dept-ABES Engg College, Ghaziabad (U.P.), India

Index Terms - Clustering algorithms, categorical dataset, numerical dataset, clustering, data mining, pattern discovery, genetic algorithm.

1.0 INTRODUCTION

The basic operation in Data Mining is partitioning of sets of objects present in the database into homogenous clusters or groups is the basic work in data mining. The beneficial way in numerous tasks likeclustering, image processing, sequence analysis, market research, pattern recognition, spatial analysis, economics etc. To implement the operation of partitioning clustering is the most widely used approach. This technique partitions the sets of objects into unsupervised classifiers in such a way that the objects contained in the common cluster are more similar to each other than objects in indifferent clusters.

Data mining and warehousing differs from other traditional applications and analysis of clusters in such a way that it deals with large high dimensional data.According to this attribute, manyunsupervised classification algorithms are discontinued to beused. One more characteristic is that data mining data often contains all types of mixed attributes in real life practical applications. The traditional method to handle categorical, nominal, ratio-scaled or ordinal attributes as numeric with the help of dissimilarity matrices (after calculating Euclidean/Manhattan/Minkowaski distances and applying normalization (standard deviation/ Z-score or min-max normalization on the results) and applying the related algorithms for numeric values, but due to the unordering of many categorical domains it does not always yield useful and meaningful results.

Many already available unsupervised classification algorithms can handle either only numeric attributes or both data types but not efficient when clustering is performed on large sets of data. Few algorithms can perform both well, such as k-prototypes etc.

We give a new cost function for clustering to process large sets of data with mixed numeric and categorical and other values by doing changes in the common used trace of within cluster dispersion matrix. In clustering process, we introduce genetic algorithm(GA) so that the cost function can be minimized. The benefit of high search efficiency is achieved in GA as GA uses search strategy globally and also implements in parallel.

The remaining paper is organized as follows: Some mathematical preliminaries of the algorithm are included in the next section. Then GA is briefly discussed with modified and efficient cost function for all the data sets. In last section there are summaries the discussions.

2.0 BETTERMENT BY THE USE OF GENETIC ALGORITHM

With the basic features of GA like encoding, crossover, mutation, appropriate fitness function and reproduction with survivor selection, the GA can be able to design better clustering and unsupervised classification operations.

The proposed approach can be described with experiments and their results. The algorithm can be run on real-life datasets to test its clustering performance against other algorithms. At the same time, its properties are also empirically studied. One observation from the above analysis is that our algorithm's computation complexity is determined by the component clustering algorithms. So far, many efficient clustering algorithms for large databases are available, it implicate that our algorithms will effective for large-scale data mining applications, too.

3.0 COST FUNCTION, INITIAL POPULATION AND SELECTION

Initial Population

The size of the initial population is an important issue because a large population can effectively sample the parameter space. However the larger the population, the higher the computational cost. A compromise must be found. An interesting means to both have an effective sampling and a reasonable computational cost is to decrease the population after the first iteration of the process. For instance, take an initial population of 1000 chromosomes, then choose the 500 better chromosomes and work with this population's size until the end.

Cost Function

The cost function represents the problem we want to solve. For instance, the cost function of the well-known traveling salesman is the distance the salesman has to cover to visit all the towns exactly once. Most search problems may be posed as the search for the optimal value of a function satisfying a set of constraints. The function represents the relationships between the different parameters which we seek to optimize. When those relationships are well-defined and simple enough to be modeled mathematically (e.g. by convex functions), the analytical methods (e.g. Lagrange multipliers) of mathematics should be applied to the problem. When those relationships are so complex as to appear unpredictable or random, the model itself may be ill-posed and only random or exhaustive search offers any hope of an answer. Genetic algorithms are well suited for the real-world problems which lie between these two extremes.

Selection

We have now a set of chromosomes. In order to enhance the average fitness of the population, we will generate a new population from the previous one according to the quality of each chromosome: the higher fitness value of a chromosome, the higher its probability to be included in the new generation. The standard way to do this is called the casino roulette method:

4.0 COST FUNCTION FOR NUMERIC DATA CLUSTERING

The trace of the within cluster dispersion matrix is the widely used cost function.The cost function is defined as:

$$C(W) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^2 (\mathrm{d}(x_j, x_i))^2 \ , w_{ij} \epsilon \{0,1\} \quad (1)$$

Here, w_{ij} is the degree of membership of x_j belonging to cluster i. **W** is a $k \times n$ order partition matrix. The function d(.) is a measure of dissimilarity often defined as the Euclidean distance. For data set having real attributes, i.e., X $\subset R^m$, we have

$$\mathrm{d}(x_j, x_i) = \left(\sum_{l=1}^{m} | x_{jl} - x_{il} | 2\right)½ \quad (2)$$

Since, w_{ij} indicates x_j belonging to cluster i, and $w_{ij} \epsilon [0,1]$, we call **W** to be a hard k-partition.

5.0 COST FUNCTION FOR MIXED DATA CLUSTERING

5.1Max-Min Normalization for numeric data

For clustering the numeric data, first we will normalize numeric data so as to prevent the dominance of particular attribute. For which the normalization formula is as follows:-

$$\mathrm{n_i} = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \times (Rh - Rl) + Rl \quad (3)$$

Where, x_i is the i-thobject.Rh and Rlare the higher and lower ranges respectively. N is the new normalized matrix containing all types of data.

5.2 Normalizing ratio-scaledvalues:-

First, we will take log of the ratio-scaled values, given as

$$f(n) = \log(n) \quad (4)$$

5.3Normalizing ordinal values:-

First we assign ranks to the values as, better the value higher the rank and vice versa. Now, based on their ranks we will normalize them. Give 1 to the highest rank and 0 to the lowest one and other ranks get the value as:

$$\prec(\mathrm{r}) = \frac{1}{no.of different ordinal values - 1} \times (\mathrm{r}\text{-}1) \quad (5)$$

5.4 Normalizing categorical values:-

If the two values match put value 1 and otherwise 0.

$$\delta(a, b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases} \quad (6)$$

6.0 RE-DEFINING COST FUNCTION

When Nhas attributes with numeric and mixed values, assuming that eachobject is denoted by n_i=[$n_{i1}^r, \ldots, n_{it}^r, n_{i,t+1}^c, \ldots\ldots, n_{im}^c, n_{i,m+1}^b, \ldots\ldots, n_{i,y}^b, n_{i,y+1}^o, \ldots \ldots,$ $n_{iu}^o, n_{i,q+1}^{rs}, \ldots \ldots, n_{is}^{rs}$,],the dissimilarity between two mixed-typeobjects $n_i$and $n_j$can be measured by the following Eq.(7)

$$d(n_i, n_j) = [\left(\sum_{l=1}^{t} |n_{il}^r - n_{jl}^r|)2 + \prec_1 . (\sum_{l=t+1}^{m} |n_{il}^c , n_{jl}^c|)2 + \prec_2 . \left(\sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b \right|)2 + \prec_3 . (\sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o \Big|\Big) 2\Big) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs}|2)]½$$

Whereall the terms are squared Euclidean distance measure on the mixed attributes.

Using Eq. (7) for mixed-type objects, we can modify the cost function of Eq. (1) for mixeddata clustering. In addition, modifythecost function to extend the hard k-partitioning as:

$$C(W) = \sum_{l=1}^{K}\left(\sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r|2 + \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c|2 + \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b |2 + \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o |2) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs} |2), \quad w_{ij} \varepsilon [0,1] \quad (8)$$

Let $C_i^r = \sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r | 2$

$$C_i^c = \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c |2 C_i^b$$

$$= \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b |2 \; C_i^o$$

$$= \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o |2$$

(9)

We rewrite Eq.(8) as:

$$C(W) = \sum_{i=1}^{k} (C_i^r + C_i^c + C_i^b + C_i^o) \qquad (10)$$

7.0 GA-BASED CLUSTERING ALGORITHM FOR MIXED DATA

Clustering is a fundamental and widely applied method in understanding and exploring a data set. Interest in clustering has increased recently due to the emergence of several new areas of applications including data mining, bioinformatics, web use data analysis, image analysis etc. To enhance the performance of clustering algorithms, Genetic Algorithms (GAs) is applied to the clustering algorithm. GAs are the best-known evolutionary techniques. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering. This paper present some existing GAbased clustering algorithms and their application to different problems and domains.

8.0 GENETIC ALGORITHM

In the field of artificial intelligence, a genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a meta heuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetics, computational,science, engine ering, economics, chemistry,manufacturing, mathematics, physi cs, pharmacometrics and other fields.

9.0 GA-BASED CLUSTERING ALGORITHM

9.1 Algorithm:

Step1. Begin
Step2. Define pop-size as desired population size
Step3.Randomly initializes pop-size population
Step4.While (Ideal best found or certain number of generations met)
O Evaluate fitness
O While (number of children=population size)
O Select parents
O Apply evolutionary operators to create children
O End while
Step5. End While
Step6. Return Best solution
Step7. End

First of all, the following three problems should be solved to employ GA:
(1) Encoding of the clustering solution into the gene string.
(2) Designing of a reasonable fitness function.
(3) Selection or designing genetic operators including their parameters that guarantees fast convergence.

9.2 Encoding: From Eq.(1) and (8), it is well known that the purpose of clustering is to obtain a (fuzzy) partition matrix **W.** Then using the fitness function (stated below) we can improve the chances of a particular data point to be chosen.Then after selecting that particular cluster we can further subdivide the data points in the cluster, based on their fitness values.
Note that since we process data having mixed attributes, in parallel to numeric parameter mixed parameters are also there in gene string. Therefore, not ordered for binary attributes and can be directly encoded rather than doing normalization first.

9.3 Fitness function: We are taking the fitness function such that fitness value is inversely proportional to the cost function value, i.e., the fuzzy clustering partition is better when the cost function is smaller. So GA asks for a larger fitness value. Hence, fitness function is defined with the use of clustering

cost function. Exponentially increased cost function will sharply reduce the fitness function.

$f(g) = \frac{1}{1+e^{C(W)}}$(11)

9.4 Genetic operators: A genetic operator is an operator used in genetic algorithms to maintain genetic diversity, known as mutation and to combine existing solutions into others, crossover. The main difference between them is that the mutation operators operate on one chromosome, that is, they are unary, while the crossover operators are binary operators. Genetic variation is a necessity for the process of evolution. Genetic operators used in genetic algorithms are analogous to those in the natural world: survival of the fittest, orselection; reproduction (crossover, also called recombination); and mutation.

Types of Operators

1. Selection (genetic algorithm)
2. Crossover (genetic algorithm)
3. Mutation (genetic algorithm) and selection probability is:

$P_s(g_{(i)}) = \frac{f(g_i)}{\sum_{i=1}^{n} f(g_i)}$ (12)

Operation probabilities for crossover and mutation are assigned as Eq. (13)

$$P_c(g_i, g_j) = \begin{cases} \frac{\alpha_1(f_{max} - f')}{f_{max} - \bar{f}} f' \geq \bar{f} \\ \alpha_2 otherwise \end{cases}$$

$$P_m(g_i) = \begin{cases} \frac{\alpha_3(f_{max} - f(g_i))}{f_{max} - \bar{f}} f(g_i) \geq \bar{f} \\ \alpha_4 otherwise \end{cases} \quad (14)$$

where, $f_{max} = max_{l=1}^{N}\{f(g_l)\}$

$\bar{f} = \sum_{l=1}^{N} f(g_l)$, $f' =, f(g_j)\}$,and$\alpha_i \epsilon [0, 1]$

Apart from the operators mentioned above, a new operator for the clustering algorithm is defined.

Gradient operator: Changes in the existing weights are done as per the formula:

It includes two steps iteration as:

$$w_{ij} = \sum_{l=1}^{k} \frac{(d(x_j, x_i))2}{(d(x_j, x_i))^2}, i, \text{j} \quad (15)$$

10.0 A REAL-LIFE PRACTICAL SAMPLE DATA TABLE OF MIXED DATA TYPES

We are representing the real life concept of our approach by taking the data of 5 employess working in a company. Here we will use every kindof data (related to all data types) to show that our method works for every kind of data.In this example :

We are taking the weighted matrix (W_{ij}) as:

Table 1

	1	2	3	4	5
1	**0**				
2	**0.4**	**0**			
3	**0.2**	**0.2**	**0**		
4	**0.1**	**0.3**	**0.2**	**0**	
5	**0.5**	**0.2**	**0.1**	**0.4**	**0**

Test-1 cantains salary of an employee (numeric data)

Test-2 shows whether the employee is male or female(binary data- Male=1/ Female=0)

Test-3 shows the department to which employee belong (categorical data)

Test-4 depicts the ability of an employee (ordinal values) Exc.-Excellent, Fair or Good

Test-5 shows avg. credit points alloted according totheir performance (ratio-scaled values)

Last Column shows the log value of ratio-scaled data type.

Table 2

Object-id	Test-1	Test-2	Test-3	Test-4	Test-5	Log
1	**25K**	**M**	**Code-A**	**Exc.**	**445**	**2.65**
2	**40K**	**F**	**Code-B**	**Fair**	**22**	**1.34**
3	**55K**	**M**	**Code-C**	**Good**	**164**	**2.21**
4	**27K**	**M**	**Code-A**	**Exc.**	**1210**	**3.08**
5	**53K**	**F**	**Code-B**	**Fair**	**38**	**1.58**

The Table 2 is converted into the normalized matrix using the above equations.(3),Eq.(4),Eq.(5),Eq.(6)

Table 3

	1	2	3	4	5
1	**0**	**0**	**0**	**0**	**0**
2	**0.5**	**1**	**1**	**1**	**1.31**
3	**1**	**0**	**1**	**0.5**	**0.44**
4	**0.0666**	**0**	**0**	**0**	**0.43**
5	**0.9333**	**1**	**1**	**1**	**1.07**

We calculate the value of the expression (stated below) to be further used in Eq. (8)

$$\sum_{j=1}^{n} d(x_j, x_i)$$

Table 4

	1	2	3	4	5
1	**0**				
2	**4.9961**	**0**			

3	**2.4436**	**2.2569**	**0**		
4	**0.1893**	**3.962**	**2.1213**	**0**	
5	**5.0159**	**0.2453**	**1.6153**	**4.1607**	**0**

Now for Eq. (8) we are calculating the value of the expression:

$$\sum_{j=1}^{n} w_{ij}^{2} \left(d(x_j, x_i)\right) 2$$

Table 5

	1	2	3	4	5
1	**0**				
2	**0.0799**	**0**			
3	**0.0977**	**0.0903**	**0**		
4	**0.0018**	**0.3566**	**0.0848**	**0**	
5	**1.2539**	**0.0098**	**0.0165**	**0.6571**	**0**

Now we will calculate the expression:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^{2} \left(d(x_j, x_i)\right) 2$$

Table 6

1	**1.3455**
2	**0.5366**
3	**0.2013**
4	**1.1063**
5	**1.9373**

Now using the Eq. (11) we find the fitness value of the above calculated values (above 5 tuples):

Table 7

1	**0.2066**
2	**0.3689**
3	**0.4498**
4	**0.2497**
5	**0.1259**

First we arrange the above values in ascending order and label each one of them and then using Eq.(12) calculate the selection probability $P_s(g_{(i)})$

Table 8

1	**0.1474**
2	**0.2633**
3	**0.3204**
4	**0.1782**
5	**0.0898**

11.0 ANALYSIS ON OUR EXPERIMENTAL RESULTS

By the above calculated tables, we can easily verify the dissimilarity matrices of our real life experimental data shown in tabular structure,

We can comfortablydecide the set of clusters based on the fitness values.We are taking the threshold value for our method to be 0.22. Data item 1 and 5whose fitness value lie below the threshold value can be grouped together in the cluster and the other three tuples can be grouped in another.

Now these clusters can be improved using GA and using the selection probability.

Result:

So there can be two clusters:

C1:- data items 1 and 5.

C2:- data items 2,3 and 4

12.0 CONCLUSION

Here, to cluster large sets of data we have presented GA and the performance can be evaluated using large data sets of data. The proposed outcomes can be used to demonstrate the importanceof the algorithms in finding structures in data.

This paper puts an emphasis on the issue that uses the GA to solve the clustering problem. Though the application is specific for the business, our approach is general purpose and could be used with a variety of mixed-type databases or spreadsheets with categorical, numeric and other data values, and temporal information. With improved metrics, artificial intelligence algorithms and decision analysis tools canyield more meaningful results and agents can make better decisions.

This approach, then, can ultimately lead to vastly improved decision-making and coordinating among business units and agents alike. If a class attribute is involved in the data, relevance analysis between the class attribute and the others (or feature selection) should be performed before training to ensure the quality of cluster analysis. Moreover, most variants of the GA use Euclidean-based distance metrics. It is interesting to investigate other possible metrics like the Manhattan distance or Cosine-correlation in the future. To faithfully preserve the topological structure of the mixed data on the trained map, we integrate distance hierarchy with GA for expressing the distance of categorical values reasonably.

13.0 ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their valuable suggestions. They would also like to thank Prof. A.K. Sinha

(Dean CRAP-ABES-Engineering College, Ghaziabad) for his involvement and valuable suggestions on soft-computing in the early stage of this paper.

REFERENCES

[1]. LI Jie, GAO Xinbo, JIAO Li-cheng, "A GA-Based Clustering Algorithm for Large Data Sets withMixedNumeric and Categorical Values",National Key Lab. of Radar Signal Processing, Xidian Univ., Xi'an 710071, China

[2]. M. R. Anderberg. Cluster Analysis for Applications. Academic Press, New York,1973.

[3]. B. Everitt. Cluster Analysis. Heinemann Educational Books Ltd., 1974.

[4]. Zhexue Huang, Michael K.Ng. A fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans. on Fuzzy Systems, 7(4): 446-452, August, 1999.

[5]. Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data Mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, the University of British Columbia, Canada, pp.1-8.

[6]. R. Krovi. Genetic Algorithm for Clustering: A Preliminary Investigation. IEEE press, Pp.504-544.

[7]. J. H. Holland. Adoption in Natural and Artificial System. Ann Arbor, MI: Univ. Mich. Press, 1975.

TECHNICAL PROFILE

Mr. Rohit Rastogi received his B.E. degree in Computer Science and Engineering from C.C.S.Univ. Meerut in 2003, the M.E. degree in Computer Science from NITTTR-Chandigarh (National Institute of Technical Teachers Training and Research-affiliated to MHRD, Govt. of India), Punjab Univ. Chandigarh in 2010.

He was Asst. Professor at IMS College, Ghaziabad in computer Sc. Dept. His research interests include Data ware Housing and Data Mining, Design Analysis of Algorithm, Theory of Computation & Formal Languages and Data Bases.

He is a Sr. Asst. Professor of CSE Dept. in ABES Engineering. College, Ghaziabad (U.P.-India), affiliated to Gautam Buddha Tech. University and Mahamaya Tech. University (earlier Uttar Pradesh Tech. University) at present and is engaged in Clustering of Mixed Variety of Data and Attributes with real life application applied by Genetic Algorithm, Pattern Recognition and Artificial Intelligence.

He has served as the technical reviewer of 7 papers in IIIrd International Conference on Computing, Communications and Informatics (IC3'2014) at GCET, Greater Noida, NOIDA, India on September, 24-27, 2014 And Worked as the reviewer for the SPICES-2015 at NIT Kerala, Kojhicode for international conf. of Signal Processing and Communication…Currently working as the reviewer in the technical reviewer committee for the INDIA-2015 is Second International Conference on Information System Design and Intelligent Applications organized by Faculty of Engineering, Technology and Management, University of Kalyani, Kalyani-741235, West Bengal, India.

Currently designated reviewer on the technical program committee for the International Conference on Computing in Mechanical Engineering (ICCME-2015) (ICCME-2015). The proceedings of ICCME'15 will be published by Springer as a special volume in the Lecture Notes in Mechanical Engineering (ISSN: 2195-4356). All accepted papers will also be archived in the SpringerLink digital Library.He is UGC-NET -2014 qualified.

He has mentored around 20 Live Projects in Digital Logic Design at Graduation level like Automatic street Light Controller, Darkness detector, Visitor counter and Car Parking system etc.

He is CSI-student Coordinator of ABES-EC CSI student Chapter and life member of ISTE.

He keeps himself engaged in various competitive events, activities, webinars, seminars, workshops, projects and various other teaching Learning forums.

He has been awarded in different categories by ABES-EC, Gzb. College management for improved teaching, significant contribution, human value promotions and long service etc.

He has authored/co-authored, participated and presented research papers in various Science and Management areas in around 40 International Journals and International conferences including prestigious IEEE and Springer and 10 national conferences including SRM Univ., Amity Univ. and Bharti Vidyapeetha etc. He has guided five ME students in their thesis work and students of UG and PG in around 100 research papers. He has developed many commercial applications and projects and supervised around 30 B.E. students at graduation level projects.

His research interests include Data ware Housing and Data Mining, Design Analysis of Algorithm, Theory of Computation & Formal Languages and Data Bases. At present, He is engaged in Clustering of Mixed Variety of Data and Attributes with real life application applied by Genetic Algorithm, Pattern Recognition and Artificial Intelligence.

Also, He is preparing some interesting algorithms on Swarm Intelligence approaches like PSO, ACO and BCO etc..

Tuning, Diagnostics & Data Preparation for Generalized Linear Models Supervised Algorithm in Data Mining Technologies

Sachin Bhaskar[1], Vijay Bahadur Singh[2] and **A. K. Nayak [3]**

Submitted in July 2013; Accepted in March, 2015

Abstract - Data mining techniques are the result of a long process of research and product development. Large amount of data are searched by the practice of Data Mining to find out the trends and patterns that go beyond simple analysis. For segmentation of data and also to evaluate the possibility of future events, complex mathematical algorithms are used here. Specific algorithm produces each Data Mining model. More than one algorithms are used to solve in best way by some Data Mining problems. Data Mining technologies can be used through Oracle. Generalized Linear Models (GLM) Algorithm is used in Regression and Classification Oracle Data Mining functions. For linear modelling, GLM is one the popular statistical techniques. For regression and binary classification, GLM is implemented by Oracle Data Mining. Row diagnostics as well as model statistics and extensive co-efficient statistics are provided by GLM. It also supports confidence bounds.. This paper outlines and produces analysis of GLM algorithm, which will guide to understand the tuning, diagnostics & data preparation process and the importance of Regression & Classification supervised Oracle Data Mining functions and it is utilized in marketing, time series prediction, financial forecasting, overall business planning, trend analysis, environmental modelling, biomedical and drug response modelling, etc.

Index Terms – GLM, Linear regression, Logistic regression, ODM,Tuning and Diagnostics for GLM

1.0 INTRODUCTION

GLM includes and extends the class of linear models[1]. The set of restrictive assumptions are made by linear models and also the most importantly the conditions are generally distributed on the value of predictors with a constant variance irrespective of predicted response values. An interpretable model form, able to compute specific diagnostic information about the quality and computational simplicity are included by the advantages of linear models and their restrictions. These restrictions are relaxed by Generalized Linear models which are not found is general practice.

[1]*Bihar Institute of Public Administration & Rural Development, Patna, Bihar, India*
[2]*L. N. Mishra Institute of Economic Development & Social Change, Baily Road, Patna, Bihar, India*
[3]*Zakir Hussain National Institute, Patna, Bihar, India.*
E-mail: [1]sachinbhaskar007@yahoomail.com,
[2]vbsinghpat@yahoo.co.in and [3]akn_iibm@yahoo.com

We find that the sum of terms in a linear model typically have large ranges encompassing very negative and positive values. For example in case of binary response, the response of probability is liked in the range [0,1]. We have two mechanism namely variance function and linear function where linear model assumptions are violated by responses which GLM accommodate. The variance function expresses the variance as a function of the predicted response thereby accommodating responses with non-constant variances like binary responses. The linear function transforms the target range to potentially -ve infinity to +ve infinity for maintaining the simple form of linear models. Two widely known members of the GLM family of models with their most popular link and variance functions included in Oracle Data Mining are as follows:

- Linear regression with the identity link and variance function equal to the constant 1 (constant variance over the range of response values). [1]
- Logistic regression with the log it link and binomial variance functions. [2]

GLM is a well established parametric modelling technique. Assumptions about the distribution of the data are made by parametric models. Parametric models become more efficient than the non-parametric models when assumptions are met. Assessing the extent to which the assumptions are met, is involved by the challenge in developing models of this type. That is why for developing quality parametric models, quality diagnostics are the key factors.

2.0 GLM INORACLE DATA MINING

2.1 Interpretability and Transparency

It is easy to interpret the Oracle Data Mining GLM models. Several diagnostics and statistics are generated by each model build. Transparency is also an important characteristic. Model details also describe key characteristics of global details providing high-level statistics and coefficients. **[3][4]**

2.2 Wide Data

To handle wide range of data, GLM is uniquely suited. The algorithm can build and score quality models that uses a virtually limitless number of predictors (attributes). The constraints imposed by the system resources are the only constraints.

2.3 Confidence Bounds

GLM is able to predict confidence bounds. GLM is able to predict confidence bounds. Apart from the predict, the best estimate and a probability, it identifies probability

(classification) and an interval wherein the prediction (regression) will lie. The width of the interval is dependable on the precision of the model and a user-specified confidence level The confidence level is a measure to know the true value that lie within a confidence interval computed by the model. 95% is the popular choice of confidence level. For Example - a model mighty predict that an employee's income is $130K and that we can be sure that around 95% sure that it lies between $85K and $150K. The value so obtained is configurable, although Oracle Data Mining supports 95% confidence by default
It is returned along with the coefficient statistics. PREDICTION_BOUNDS SQL function is to obtain the confidence bounds of a model prediction can also be used

2.4 Ridge Regression
The best regression models are the predictors which correlate highly with the target but there is very little correlation between the predictors themselves. Multi-collinearity is used to describe multivariate regression with correlated predictors. Multi-collinearity is compensated by the technique called Ridge regression. Ridge regression is supported by Oracle Data Mining for both classification and regression mining functions. If the singularity (exact multi-collinearity) in the data is found, ridge is automatically used by algorithm. Information about the singularity is returned in the global mode[6][7].

2.5 Build Settings for Ridge Regression
We can choose to explicitly enable ridge regression by specifying the GLMS_RIDGE_REGRESSION setting. If we exclusively enable ridge, we can use the system-generated ridge parameter or we can supply our own. Explicitly enable ridge regression by specifying the GLMS_RIDGE_REGRESSION setting can be choosen. If ridge explicitly enabled, the system-generated ridge parameter can be used or we can supply our own. The ridge parameter is also calculated automatically in case if the ridge is used automatically.
The build settings for ridge can be summarized as[5]:
• GLMS_RIDGE_REGRESSION — Whether or not to override the automatic choice made by the algorithm regarding ridge regression.
• GLMS_RIDGE_VALUE — The value of the ridge parameter, used only if you specifically enable ridge regression.
• GLMS_VIF_FOR_RIDGE — Whether or not to produce Variance Inflation Factor (VIF) statistics when ridge is being used for linear regression.

2.6 Ridge, Confidence Bounds, Variance Inflation Factor for Linear Regression
Models built with ridge regression do not support confidence bounds[8]. Variance Inflation Factor (VIF) statistics for linear regression models are produced by GLM, unless they were built with ridge. VIF with ridge by specifying the GLMS_VIF_FOR_RIDGE setting can be exclusively requested. VIF with ridge will be produced by the algorithm only in case if enough system resources are available.

2.7 Ridge and Data Preparation
Different data preparations are likely to be produced different results in terms of model coefficients and diagnostics in case the ridge regression is enabled. Automatic Data Preparation for GLM models, especially when ridge regression is used, be enabled; a Oracle Corporation recommends[9].

3.0 TUNING AND DIAGNOSTICS FOR GLM
A number of model builds involved in the process of developing a GLM model. To evaluate and determine the quality of model, each build generates many statistics. To change the model settings or making other modifications can be tried on the basis of these diagnostics.

3.1 Build Settings
Build settings is basically used for the purpose of specification of:

- Coefficient confidence: The default confidence which is used widely is 0.95. The degree of certainty that the true coefficient lies within the confidence bounds computed by the model, is indicated by GLMS_CONF_LEVEL setting.
- Row weights: This checks whether a column is containing a weighting factor for the rows or not and the situation is indicated by - ODMS_ROW_WEIGHT_COLUMN_NAME setting.
- Row diagnostics: This is used to identifya table to contain row-level diagnostics. This is indicated by GLMS_DIAGNOSTICS_TABLE_NAMEsetting.

There are additional build setting which are used for:

- Controlling the utilization of ridge regression[10].
- Handling procedure's specification for missing values which are not present in the training data[9].
- Specification of target values to be used as reference in logistic regression model[7].

3.2 Diagnostics
To evaluate the quality of the model GLM models generate many metrics to help us.

3.3 Coefficient Statistics
Both linear and logistic regression return the same set of statistics but statistics that do not apply to the mining function are returned as NULL [6][7]. The GET_MODEL_DETAILS_GLM function in DBMS_DATA_MININis used for the purpose of returning the coefficient statistics.

3.4 Global Model Statistics
For linear and logistic regression a whole new method of separation is adopted by returning separate high-level statistics which describes the model as a whole[6][7]. Only fewer global

details are returned when the ridge regression is enabled and this makes the adjacent procedures convenient[10].
The GET_MODEL_DETAILS_GLOBAL function in DBMS_DATA_MINING returns global statistics.

3.5 Row Diagnostics

By specifying the name of a diagnostics table in the build settingGLMS_DIAGNOSTICS_TABLE_NAME,
configuration of GLM models could be done to generate per-row statistics[6][7].
Row diagnostics is generated by a case ID which is required by GLM. An exception is raised in the process in casewe provide the name of a diagnostic table but the data does not include a case ID column.

4. 0 DATA PREPARATION FOR GLM

For both linear and logistic regression Automatic Data Preparation (ADP) implements suitable data transformation [11].

4.1 Data Preparation for Linear Regression

The build data are standardized by using a widely used correlation transformation, when ADP is enabled [12]. From the attribute values for each observation the data are first centred by subtracting the attribute means. In an observation by the square root of the sum of squares per attribute across all observations, the data are scaled by dividing each attribute value. For both numeric and categorical attributes, this transformation is used.
Before standardization, When N is the attribute cardinality, categorical attributes are exploded into N-1 columns. During the explosion transformation, the most frequent value (mode) is omitted. The first value in the list is omitted during the explosion and the attribute values are sorted alpha-numerically in ascending order in case of highest frequency ties. Where ADP is enabled or not explosion transformation lies.
The described transformations (explosion followed by standardization) can increase the build data size because the resulting data representation is dense, in case of high cardinality categorical attributes. An alternative approach needs to be used to reduce disk space, memory and processing requirements. Categorical attributes are not standardized for large datasets where the estimated internal dense representation would require approx more than 1Gb of disk space. The VIF statistic should be used with caution under aforesaid circumstances [10][11].

4.2 Data Preparation for Logistic Regression

Categorical attributes are exploded into N-1 columns where N is the attribute cardinality. The Explosion transformation eliminates the most frequent value (mode). The attribute values are sorted alpha-numerically in ascending order in the case of highest frequency ties, and the first value on the list is discarded during the explosion. Explosion transformation takes place irrespective of enabling the ADP and it doesn't depend upon its mode.
Numerical attributes are standardized when ADP is enabled mode. Measure of attribute variability plays a pivotal role in scaling the attribute values and this mechanism helps in the process of standardisation. Computation of the particular measure of variability is done with respect to the standard deviation per attribute with respect to the origin (not the mean)[13].

4.3 Missing Values

In case of applying or building a model, missing values of numerical attributes with the mean and missing values of categorical attributes with the mode are automatically replaced by Oracle Data Mining.
A GLM model can be configured to override the default treatment of missing values. The algorithm can be caused to delete rows in the training data that have missing values instead of replacing them with the mean or the mode, with the ODMS_MISSING_VALUE_TREATMENT setting. However, when we apply the model, Oracle Data Mining performs the usual mean/mode missing value replacement.
As a result, we see that statistics generated from scoring often not match the statistics generated from building the model.
The transformation must be performed explicitly if we want to delete rows with missing values in the scoring the model. The rows with NULLs from the scoring data must be removed before performing the apply operation for making build and apply statistics match.
This can be done by creating a view like

```
CREATE VIEW view_name  AS SELECT * from table_name
   WHERE column_name1 is NOT NULL
   AND   column_name2 is NOT NULL
   AND   column_name3 is NOT NULL.....
```

5.0 CONCLUSION AND FUTURE SCOPE

This paper is about Generalised Linear Models (GLM) Supervised algorithm in data mining technologies specifically in terms of tuning, diagnostics and data preparation process. In today's world GLM is apopular statistical technique, especially for linear modelling since it possesses highly remarkable features with respect to present requirements. Oracle Data Mining implements GLM for binary Classificationand for Regression function. This paper emphasises on explanation of the previous work, which hasbeen reviewed for the understanding of research in the area of Data Mining technologies.

REFERENCES

[1]. Linear Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/regress.htm#CIHJIFEG
[2]. Logistic Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CIACAIFC
[3]. Tuning and Diagnostics for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABBGAHD

[4]. Transparency for GLM, http://download.oracle.com/docs/cd/B28359_01/dAtamine.111/b28129/xform_data.htm#CIAICHGH
[5]. Oracle Database SQL Language Reference, http://download.oracle.com/docs/cd/B28359_01/sErver.111/b28286/functions121.htm#SQLRF20020
[6]. Global Model Statistics for Linear Regression,http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABBIADB
[7]. Global Model Statistics for Logistic Regression, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CHDEBJEB
[8]. Confidence Bounds for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABDBIII
[9]. Data Preparation for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#CACCHJDC
[10]. Ridge Regression for GLM, http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/algo_glm.htm#BABHBBBA
[11]. Automatic and Embedded Data Preparation,http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/xform_data.htm#BABGADFF
[12]. Neter, J., Wasserman, W., and Kutner, M.H., "Applied Statistical Models", Richard D. Irwin, Inc., Burr Ridge, IL, 1990.
[13]. Marquardt, D.W., "A Critique of Some Ridge Regression Methods: Comment", Journal of the American Statistical Association, Vol. 75, No. 369 , 1980, pp. 87-91
[14]. Alex Berson& Stephen J. Smith, "Data Warehousing Data Mining and OLAP", Tata McGraw-Hill Publishing Company Limited, New Delhi.
[15]. Bhaskar Sachin, Dissertation "Managin Data Mining Technologies in Organizations: Techniques and Applications", submitted to Periyar University, Salem, for M.Phil in Computer Science, November, 2007.

Clustering of Web Usage Data using Hybrid K-means and PACT Algorithms

T. Vijaya Kumar[1] and **H. S. Guruprasad**[2]

Submitted in May 2014; Accepted in March, 2015

Abstract-Clustering is an exploratory technique that structures the data items into groups based on their similarity or relativeness. Clustering is used in Web usage scenario to form clusters of users showing same behaviour and clusters of pages with similar or related information. The Clustering of users results in the establishment of groups of users with related browsing patterns. The most popular technique to find the clusters is the K-means clustering algorithm. This paper presents a technique to improve the Web session's cluster quality using Subtractive clustering algorithm. In this paper, the Web Session clusters are obtained by using K-means algorithm initialized by subtractive clustering algorithm. The clusters formed are analysed using Profile Aggregation Based on Clustering of Transactions [PACT] algorithm. Web navigational data of the users accessing theWebsitehttp://www.enggresources.com in combined log format taken for a time window of 25 days is used as the raw data for the overall study and analysis.

Index Terms - K-means, Vector matrix, Subtractive clustering, and PACT algorithm.

1.0 INTRODUCTION

World Wide Web has grown into a very powerful and interactive media for the communication of information. Different users geographically located at different places need to access the dissimilar data types efficiently. The navigations of users with web sites generate a huge repository called Web access log file which can be analysed to discover the navigational patterns of the users. The analysis of Web access log file is termed as Web Usage Data mining. Similar to every data mining technique, Web Usage Data mining comprises of three main tasks such as web access logpre-processing, discovery of clusters followed by analysis of discovered clusters. In the pre-processing phase, the unwanted data that are not required for the next phase are removed followed by separation of users and sessions. Cluster discovery phase finds groups of similar pages or users with the similar behavioural patterns. The disseminated information on Web, results in a huge number of links for a search query. There is a need to properly organize these search results. Some search engines cluster these results and present them in an improved manner to the user. The uninterested patterns are filtered out from the user clusters and page clusters in the Cluster analysis phase. The main goal of the Clustering technique is to group together set of items which are similar to each other into the same cluster and dissimilar objects into different clusters. K-means algorithm is used by many researchers to form clusters. K-means is an algorithm with no predefined cluster centroids and non– deterministic in nature. Web access log data mining is a process of drawing out valuable information from Web access log file. The main objective of Web usage data mining is to collect the data, prototype the data as a model to represent the data, analyse the formulated model and visualize the navigational patterns of users. In this paper we have suggested an approach to obtain and analyse the clusters using hybrid K-means and PACT algorithms. First in the pre-processing phase, Server log file is given as the input and the sessions are constructed using conceptual dependency between pages and Web site structure link information which is considered as Web site graph. These identified sessions are represented using an intermediate representation called page-view matrix. Then in the Cluster discovery phase the session clusters are obtained from page-view matrix by using K-means algorithm initialized by subtractive clustering algorithm. Then these found clusters are analysed using PACT algorithm. A brief description about the review of literature is presented in Section 2. The overall architecture of K-means algorithm initialized by subtractive clustering algorithm and the details of PACT algorithm is given in Section 3. The clusters are formed as results and analysed in section 4. Conclusion for our work is briefed in section 5.

[1]*Research Scholar, Department of IS&E, BMSCE, Bangalore, Karnataka, India.*
Corresponding author.
[2]*Professor and Head, Department of CS&E, BMSCE, Bangalore, Karnataka, India.*
E-mail:[1]*vijaykrte@gmail.com and* [2]*hs_gurup@yahoo.com*

2. 0 LITERATURE REVIEW

Recently the application of data mining and artificial neural network techniques to Web log data has fascinated many researchers and they have contributed numerous procedures, tools for Web Usage mining to analyse Web navigation data. Numeral methods have been used to create models of Web navigational data using data mining and artificial neural network approaches. Various Models have been designed based on clustering algorithms, classification techniques, sequential analysis, and Markov models for discovering the knowledge from Web access log data. Web usage data clustering is the process of grouping Web users or Web sessions into clusters so

that users exhibiting the similar navigation behaviour in the same group and dissimilar navigational behaviour in different groups. K-means is deliberated as one of the main algorithms extensively used in clustering. The main advantage of the K-means algorithm is its speed and efficiency compared to other clustering algorithms. Some major drawbacks of K-means algorithm are the number of required cluster centroids must be defined before applying clustering and the random choice of initial cluster centroids. The output of the Clustering technique depends on the random choice of original cluster centroids and different runs may produce different results. In [1] the drawbacks of the standard K-means algorithm, such as the need to compute the distance from each data items to all cluster centroids, is eliminated by introducing two simple data structures. One data structure is used to hold the label of cluster and the other data structure is used to store the distance from every data item to the nearest cluster obtained in each step that can be used to find the distance in the next step. The main downside of K-means is to decide the number of clusters and initializing the centroids for the first iteration. Bashar Al-Shboul et al. have proposed an algorithm which uses genetic algorithms to initialize K-means algorithm [2]. Adaptive Resonance Theory 2 (ART2) neural network is combined with genetic K-means algorithm (GKA) to design a procedure which finds the solution for e- commerce navigational paths [3]. This technique is compared with ART2 followed by K-means and found to be better. The details of clustering algorithms and useful research directions in clustering such as semi-supervised clustering, simultaneous feature selection during data clustering, etc. are provided in [4].Fuzzy clustering [5], also known as soft clustering groups data elements that can be in more than one cluster, and a membership value is associated with each element which will result in the formation of overlapping clusters. The Fuzzy c-means algorithm is initialized by using Subtractive clustering method and experimental results showed that the modified algorithm can decrease the time complexity by reducing the number of iterations, and results in more stable and higher precision classification [6].In [7], an extension for the subtractive clustering algorithm is presented by computing the data point mountain vale. Rather than using conventional method, a kernel – induced distance measure is used in the approach. A model for cluster similar sessions by grouping the similarity matrix using Agglomerative Clustering method is presented in [8]. The session similarity is found by aligning sequences using dynamic programming. A technique for mining Web usage profiles based on subtractive clustering that scales to huge datasets is proposed in [9]. Unlike the clustering based on user description of any input parameter, they have searched in the cluster space for the finest clustering of the given Web access data. Experimental results show that the approach mines the anticipated user profiles much faster than present techniques. A new framework has been suggested in [10] using genetic algorithm and K-means clustering algorithm to improve the cluster quality of Web sessions. Costantinos Etanalyse [11] have built a model for predicting Web page by considering Web access data and Web content with weighted suffix trees. A similarity matrix is considered in the pre-processing procedure by considering the local and global sequence alignment. They have utilized the page content to enhance the proposed scheme. Fuzzy ART neural network is used to enhance the performance of the K-means in [12]. Fuzzy ART neural network technique is used to generate an initial seed value and K-means is applied as the finishing clustering algorithm. Dempster-Shafer's theory which uses evidence or beliefs from dissimilar sources is used to group users into different clusters and generate common user summaries [13]. In [14], Esin Saka et al. have presented a scheme by combining Spherical K-means algorithm and flock of agent based FClust algorithm. Spherical K-means algorithm is mainly used for clustering sparse and high dimensional data. FClust is mainly applicable for representing high dimensional data in a visualization plane. In [15],BamshadMobahser et al. have obtained aggregate usage profiles from the discovered pattern to provide effective recommendation systems for real time Web personalization systems. In [16], Parul Gupta et al. have presented a clustering technique which forms clusters from the set of documents. Every document is assigned with an identifier, so that closer document identifiers are assigned to similar documents. They have proposed an improvement for this clustering algorithm to form super clusters from mega clusters which are formed using similar clusters in a hierarchical clustering process. Their work describes the search process optimization. In [17], Naveen Aggarwal et al. have discussed on the problem of bridging the "semantic gap" between a user's need for meaningful retrieval and the current technology for computational analysis and description of the media content for Integrated Multimedia Repositories. A conceptual framework for agent-based Service Oriented Architecture (SOA) is proposed in [18], which is designed to integrate Service Oriented Architecture with the agent technology & other tactical technologies. In [19], Anil Kumar Pandey et al. have presented mutually exclusive Maximal Frequent Item set discovery based K-Means approach for finding expertise in chosen area of research. Kate A Smith [20], developed LOGSOM to represent Web pages as a two dimensional map using well known Kohonen's Self Organizing Map (SOM). The Web pages are grouped based on the interest of the Web users rather than the content of the Web page. They have considered a transaction group consisting of 235 URLs and treated them as a 235-dimensional vector as input and clustered into K = 9 clusters using K-means algorithm. For SOM output they have considered a 16 X 16 map of 256 nodes.

3.0 SYSTEM DESIGN

The main objective of the proposed system is to cluster the Web usage data using hybrid clustering and analyse the clusters using PACT algorithm. Fig. 1 depicts the overall architecture of the proposed model. We have considered the Webusage navigational data taken for a time window of 25 days of the Website ***http://www.enggresources.com*** for experimental study and discussion. In the Pre-processing Phase Data cleaning, Users Identification and Session Identification are considered

to obtain distinct users and sessions. In Data cleaning the raw server log file is cleaned and only relevant data is taken for further cluster discovery and analysis. Combinations of IP address & user agent are used to identify distinct users. In the next phase Session construction is done based on the time heuristic and navigation approach along with concept hierarchy. Session identification considers all pages accessed by single user and splits all pages into sessions. A sequence of requests made by a single user with a unique IP address on a particular Web domain for a pre-defined period of time is considered as a session. There are several approaches to construct sessions. In time heuristic, if the time spent on a page exceeds a certain threshold, or if the time between two page requests go beyond a threshold time limit then it is assumed

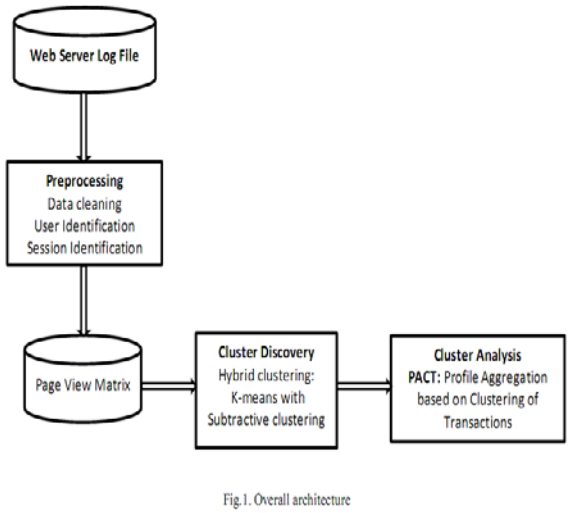


Fig.1. Overall architecture

Figure 1: Overall architecture

that a new session has been created. We have used concept switching and navigation approach with timeout as a criterion for creating user sessions [21]. Then these sessions are represented as click stream matrix. Click stream matrix can be formed by placing the session v/s sequence of pages visited in the respective session. Click stream matrix describes relationship between Web pages and sessions. To form this matrix, first we need to index each unique entry in log file and then form the matrix by placing the sequence of page visited against each session. Click stream matrix is then converted into numerical format called page view matrix. Page view matrix is constructed by building a page set of size n as pages$\{p_1, p_2, p_3, \dots\dots\dots, p_n\}$and user session set of size m as $\{ s_1, s_2, s_3, \dots\dots\dots, s_m\}$ and corresponding entry in the matrix is considered as weight for the page, it can be calculated by number of hits to the page multiplied by page hit weight which consider has 0.01.

Weight of the page $P_{i,j}$ = {frequency of access to page p_j in sessionss_i} * Page hit weight

In the Cluster discovery phase clusters are obtained by K-means clustering algorithm initialized by subtractive clustering algorithm, which takes page view matrix as input and produces the optimal number of clusters as output. K-means algorithm takes page-view matrix and number of clusters as input and it mark each session with cluster it belongs to. The modified K-means algorithm for the Web usage domain can be summarized as follows.

Step1. Consider the data set in whish sessions are represented as page-view matrix and select K points which characterize initial group centroids.

Step2. Calculate the distance between each session and every centroid and assign the session to the centroid cluster with the minimum distance.

Step3. When all sessions have been assigned to clusters, the centroid positions are calculated again by considering the cluster data points.

Step4. Repeat Steps 2 and 3 until there is no change in the centroid positions.

One of the requirements of K-means algorithm is to specify the number of centroids K before the algorithm is applied. It is difficult to guess the number of centroids for a given data set. We have used Subtractive clustering for approximating the number of centroids and the cluster centres in a dataset. The subtractive clustering technique assumes that each data point can be a promising cluster centre. Any data item which has more data items in its vicinity will have more chance of becoming a cluster centroid than data items which have less data items in its neighbourhood. Based on this principle, the potential value for each data item is computed by the following formula:

$$P_i = \sum_{j=1}^{n} e^{-4\|x_i - x_j\|^2/R_a^2}$$

Where x_i, x_j are data items and R_a is a constant value defining the range of the vicinity. The potential of the remaining data itemsx_i, is then revised by

$$P_i \Rightarrow P_i - P_k^* e^{-4\|x_i - x_k\|^2/R_b^2}$$

whereR_b is a positive constant (R_b>R_a). Thus, the data items near the first cluster centre will have greatly condensed potential value, and therefore will have very less chance of to be getting selected as the next cluster centroid. The constant R_b is the radius defining the vicinity that has a lesser potential value thanR_a. The value of R_b is set to be greater thanR_a to avoid getting nearby cluster centres. This process continues until no new cluster centroid is found. The number of clusters and the cluster centroids along with the page-view matrix is given as the input to K-means algorithm to obtain clusters. Then the PACT algorithm is used to analyse these obtained clusters to produce the aggregate profile for each Web transaction cluster. For each cluster we compute the mean vector. The measurement value for each page-view in the mean vector is the ratio between total page-view weights of all transactions and the total transactions in the cluster. The importance of any page p in a cluster is provided by its mean

vector measurement value. Page-views in the mean vector can be sorted according to these measurement values and lower measurement value page-views can be filtered out to obtain set of page view-value pairs that can be used to characterize the group of users showing similar navigational behaviour as aggregate usage summaries. These summaries can be used by the recommendation engines to provide the recommendation.

We can build the aggregate usage summary pr_{cl}, for any cluster *cl*, as a page-view weight pair by computing the mean vector of *cl* using the following formula.

$$pr_{cl} = \{(p, weight(p, pr_{cl})) \mid weight(p, pr_{cl}) \geq \mu\}$$

Where,

weight (p, pr_{cl}), of the page p within the aggregate usage summary$pr_{cl}$is given by

$$weight(p, pr_{cl}) = \frac{1}{|cl|} \sum_{s \epsilon cl} w(p, s);$$

| *cl* | is the number of sessions in cluster *cl*

$w(p, s)$is the weight of page p in session vector s

An outline of the Hybrid clustering and PACT algorithms for our system is summarized below.

Input: Sessions constructed from the pre-processing phase. Each requested URL is assigned with a unique number.

Output: Web user clusters and Recommendations

Step1. A set of m sessions are constructed from user transactions consisting of subset of n Web pages $\{p_1, p_2, p_3, \ldots\ldots\ldots, p_n\}$. These sessions are converted into page-view matrix.

Step2. Use subtractive clustering algorithm to approximate the optimal numbers of clusters and cluster centroids.

Step3. Cluster the data items usingK-means algorithm.

Step4. Obtain the recommendations using PACT algorithm.

4. Experimental Design and Results

For experimental study and analysis, we have considered the Web usage navigation data from access log files of the Web site ***http://www.enggresources.com*** collected for a time window of 25 days. Concept based Website graph is constructed as an additional input using concept hierarchy and Web site link information. We have used a tool called Web log Filter to remove fields which are not required for further analysis from access log files such as error records, requests for images and multimedia data. For further processing the important fields like IP address of the user who accessed the Web site, timestamp which represents the data and time of access, user agent details like browser information, request page and the page from the request is made called referrer are retained. User separation is considered as the next step in the pre-processing phase. In user separation, IP address and user agent are used to determine the users. In session construction, we have combined two trivial approaches, such as Time based approach and Navigation based approach along with concept name match approach for identifying user sessions. Then these sessions are represented as click stream matrix and later converted this into page view matrix. In Cluster Discovery phase we have obtained clusters by using K-means algorithm initialized by subtractive clustering algorithm. One of the major problems with K-means is to determine the optimal number of clusters and its centres which is done randomly. Subtractive clustering is used to approximately finding the number of clusters and the cluster centroids in a set of data. The subtractive clustering method assumes that each data item can be considered as cluster centre. A data item with more data items in the vicinity will have a higher chance to become a cluster centroid than data points with fewer data items in the vicinity of the data point. PACT algorithm is used to analyse these obtained clusters. The mean vector for each cluster is computed and the measurement value for each page view in the mean vector is computed by taking the ratio between the total page view weights of all transactions and the total transactions in the cluster. Clusters obtained by using K-means algorithm is plotted in Fig 2.The major drawback with K-means algorithm is determining the optimum number of clusters and its centres to initialize the K-means. The hybrid K-means algorithm uses the subtractive clustering algorithm to initialize the K-Means algorithm by providing the optimal number of clusters and its centres. The potential value for each data item is calculated based on the density of nearby data points. K-means initialized by using subtractive clustering algorithm is termed as hybrid K –means clustering and clusters obtained by using hybrid K-means is plotted in Fig 3.

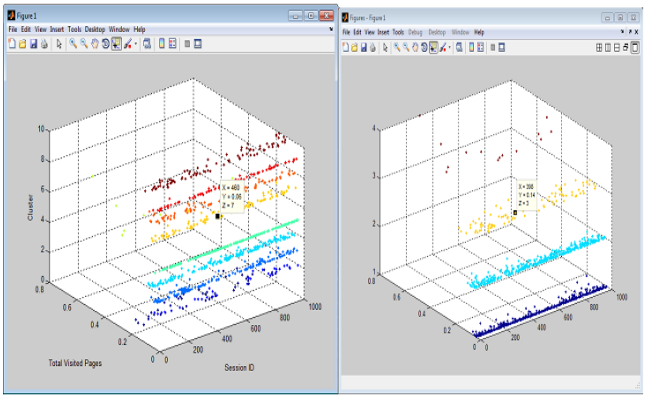


Fig. 2. Clusters obtained using K-means

Fig. 3. Clusters obtained using hybrid K-means

Figure 2: Clusters obtained using K-means

Figure 3: Clusters obtained using hybrid K-means

Fig4 depicts the Cluster-1 user segment interest. The total number of session in the cluster1 is "2286" and threshold considered is "page weight = 0.001". From Fig 4 we can observe that given a new user who shows interest in "Page32", "Page33" and "Page34", this pattern may be used to conclude that the pages "Page28", "Page3" and "Page31"may be recommended to this user. Fig 5 depicts the Cluster-2 user segment interest. The total number of session in the cluster2 is "1812" and threshold considered is "page weight = 0.0002". From Fig 5 we can observe that given a new user who shows interest in "Page28", "Page34" and "Page33", recommendation engine can consider this pattern to conclude that the recommendation engine might recommend any one of the other

pages in the above list to the user based on the order of their weight. Similar results are depicted for Cluster-3 and Cluster-4 user segments with "page weight = 0.0005" in Fig 6 and Fig 7 respectively. From Fig 6 we can observe that given a new user who shows interest in "Page31", "Page32" and "Page33", recommendation engine can consider this pattern to conclude that the recommendation engine might recommend any one of the other pages in the above list to that user based on the order of their weight. From Fig 7 we can observe that given a new user who shows interest in "Page22", "Page23" and"Page27",recommendation engine can consider this pattern

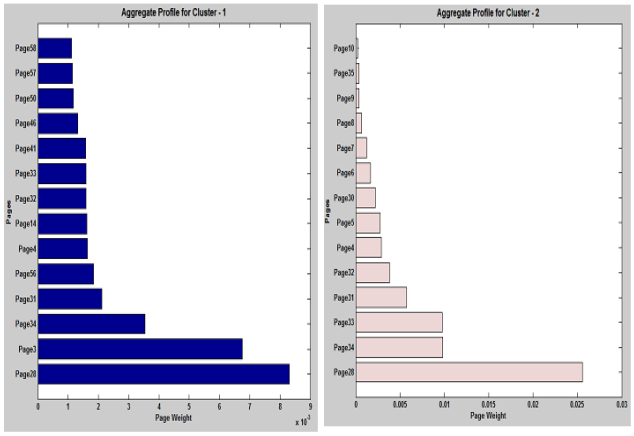


Fig. 4. Cluster-1 user segment interest

Fig. 5. Cluster-2 user segment interest

Figure 4: Cluster-1 user segment interest
Figure 5: Cluster-2 user segment interest

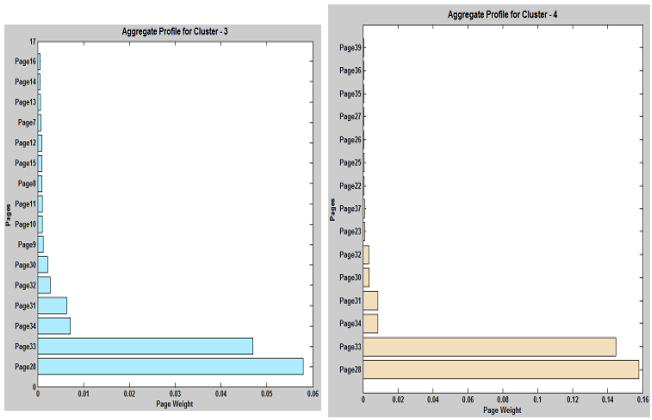


Fig. 6. Cluster-3 user segment interest

Fig. 7. Cluster-4 user segment interest

Figure 6: Cluster-3 user segment interest
Figure 7:Cluster-4 user segment interest

to conclude that the user might belong to this segment and recommendation engine might recommend any one of the other pages in the above list to that user based on the order of their weight.

5.0 CONCLUSIONS

Clustering and Analysis approach for Web usage data using hybrid K-means clustering algorithm and PACT algorithm is presented in our proposed scheme. The sessions for clustering phase are obtained by using conceptual dependency between pages and Website structure link information which is considered as Web site graph. Then these sessions are represented as click stream matrix and later converted this into page view matrix. Then clusters are formed by using K-means clustering algorithm initialized by subtractive clustering algorithm. Then clusters are analysed by using PACT algorithm to give the recommendations. As a future work, we can improve this as a recommendation engine to compare current request with navigation pattern in each cluster and come up with the recommendations.

REFERENCES

[1]. Shi Na, Guan yong, and Liu Xumin, "Research on K-means clustering algorithm" Third International Symposium on Intelligent Information Technology and Security Informatics, 2010 IEEE.

[2]. Bashar Al-Shboul and Sung-HyonMyaeng, "Initializing K-means using Genetic Algorithms" World Academy of Science, Engineering and Technology, 54 2009.

[3]. R. J. Kuo, J. L. Liao and C. Tu, "Integration of ART2 neural network and genetic K-means algorithm for analysing Web browsing paths in Electronic commerce", Decision Support Systems 40 (2005) 355-374, www.sciencedirect.com.

[4]. Anil K. Jain, "Data clustering: 50 years beyondK-means", Pattern Recognition Letters 31 (2010) 651-666.Journalhomepage www.elsevier.com/locate/patrec.

[5]. Zahid Ansari, A Vinay Babu,WaseemAhamed and Mohammed FazleAzeem, "A Fuzzy set theoretic approach to discover user sessions from Web navigational data", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE.

[6]. Qing yang, Dongxu Zhang, and Feng Tian, "Aninitialization method for Fuzzy c –means algorithm usingsubtractive clustering" 2010 Third international conference on Intelligent Networks and Intelligent Systems.

[7]. Dae-Won Kim, KiYoung Lee, Doheon Lee and Kwang H. Lee, "A Kernel based subtracting clustering algorithm",Pattern Recognition Letters, 26 (2005) 879-891, www.elseveir.com/locate/patree.

[8]. Dr. K. Duraiswamy, and V. ValliMayil, "Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic Programming",Computer and Information Science Vol. 1, N0. 3, August 2008, www.ccsenet.org/journal.html

[9]. Bhushan Shankar Suryavanshi, NematollaahShiri, and Sudhir P. Mudur, "An Efficient Technique for Mining Usage profiles Using Relational Fuzzy Subtractive Clustering", Proceedings of 2005 International

Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)

[10]. N. Sujatha, and K. Iyakutty, "Refinement of Web usage data clustering form K-means with genetic algorithm", European Journal of Scientific Research ISSN 1450-216X Volume 42 Number 3 (2010), Pages.478-490.

[11]. Costantinos Dimopoulos, Christos Makris, YannisPanagis, EvangelosTheodoridis and Athanasios Tsakalidis, "A Web page usage prediction scheme using sequence indexing and Clustering techniques", Data and Knowledge Engineering 69 (2010) 371-382, www.elsevier.com/locate/datak

[12]. Sungjune Park, Nallan C. Suresh, and Bong KeunJeong, "Sequence based clustering for Web usage mining: A new experimental framework and ANN- enhanced K-means algorithm", Elsevier Data and Knowledge Engineering 65 (2008) 512 – 543.

[13]. YunjuanXie and Vir V. Phoha, "Web user clustering from Access log using Belief Function", K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada.

[14]. Esin Saka, and OlfaNasraoui, "Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence", 20th IEEE International Conference on Tools with Artificial Intelligence.

[15]. BamshadMobasher, Honghua Dai, Tao Luo, and Miki Nakaguva, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization", Data mining and Knowledge Discovery 6, 61-82, 2002.

[16]. Parul Gupta and A.K. Sharma, "A Framework for Hierarchical Clustering Based Indexing in Search Engines", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2011; Vol. 3 No. 2; ISSN 0973 – 5658.

[17]. Naveen Aggarwal, Dr.Nupur Prakash and Dr. Sanjeev Sofat, "Mining Techniques for Integrated Multimedia Repositories: A Review", BIJIT - BVICAM's International Journal of Information Technology, January – June, 2009; Vol. 1 No. 1; ISSN 0973 – 5658.

[18]. O. P. Rishi, "Service Oriented Architecture for Business Dynamics: An Agent Based Business Modelling Approach", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2009; Vol. 1 No. 2; ISSN 0973 – 5658.

[19]. Anil Kumar Pandey and T. Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2009; Vol. 1 No. 2; ISSN 0973 – 5658.

[20]. Kate A Smith and Alan Ng, "Web page clustering using a Self-Organizing Map of user navigation patterns", Decision Support Systems 35(2003) 245-256, www.elsevier.com/locate/dsw.

[21]. T. Vijaya Kumar, Dr. H.S. Guruprasad, Bharath Kumar K.M, Irfan Baig and KiranBabu S, "A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph", IJCEE, ISSN: 1793-8198 (Online Version);1793-8163(print version).

Implementation of Enhanced Apriori Algorithm with Map Reduce for Optimizing Big Data

Sunil Kumar Khatri[1] and **Diksha Deo**[2]

Submitted in June 2014; Accepted in May, 2015

Abstract — Nowadays as a result of speedy increase in data technology. Massive scale processing may be a major purpose of advanced technology. To handle with this advance progress in information assortment and storage technologies, designing, and implementation massive scale algorithms for data processing is gaining quality and big interest. In data processing domain, association rule classification and learning may be a common and well researched methodology for locating fascinating relations between variables in massive databases. Apriori is that the key algorithmic rule to get the frequent item sets. Analyzing frequent item sets may be a crucial step to find rules and association between them. This stands as a primary foundation to monitored learning, which incorporates classifier and have extraction strategies. Enforcing this algorithmic rule is crucial to infer the behavior of structured information. In scientific domain, most of the structured information in are voluminous. Process such reasonably Brobdingnagian information needs special and dedicated computing machines. fitting such associate degree infrastructure is troublesome and dearly-won. Association rule mining needs massive computation and I/O traffic capability.

This paper majorly focuses on making association rules and Map/Reduce style and implementation of Apriori for structured information. Optimize Apriori algorithmic rule to scale back communication value. This paper aims to extract frequent patterns among set of things within the dealing info or different repositories. Apriori algorithmic rule contains a nice influence for locating frequent item sets victimization candidate generation. Apache Hadoop Map cut back is employed to create the cluster. It operating relies on Map cut back programming modal. it's accustomed improve the potency and process of enormous scale information on high performance cluster. It additionally processes Brobdingnagian information sets in parallel on massive cluster of pc nodes. It provides reliable, ascendable, distributed computing.

Index Terms— Big Data, Map Reduce, Apriori Algorithm, Optimization.

1.0 INTRODUCTION

Big data is a big thing and growing rapidly and a significant interest in new analytics used in big data. To describe big data we can say that it comprises of transactions, interactions and observations. Big data introduces large volumes of unstructured data. This data changes is highly dynamic and therefore needs to be ingested quickly for analysis.

Big knowledge could be a massive factor and growing speedily and significant interest in new analytics employed in big knowledge. To explain massive knowledge we are able to say that it contains of transactions, interactions and observations. Massive knowledge introduces massive volumes of unstructured knowledge. This knowledge changes is extremely dynamic and so must be eaten quickly for analysis.

In several applications of the important world, generated knowledge is of nice concern to the neutral because it delivers purposeful data that assists in creating prophetic analysis. This data helps in modifying sure call parameters of the applying that changes the outcome of a business method.

The volume of knowledge, conjointly referred to as data-sets, generated by the application is terribly massive. So, there is a requirement of process massive data-sets expeditiously. The knowledge-set collected is also from heterogeneous sources and will be structured or unstructured data. Process such knowledge generates helpful patterns from that data will be extracted. The only approach is to use this model and insert headings and text into it as acceptable.

There are totally different layers of massive knowledge hierarchy:

1. Variety of data
   - Structured
   - semi-structured
   - unstructured
   - complex
2. Sort of data: XML datasets.
3. Rate of data: Real time, close to real time, each minute, every hour, daily.
4. Volume of data: computer memory unit, Petabytes, Terabyte.
5. Sorts of analysis: Classification analysis, pattern recognition, regression, text-mining, clustering, anomaly detection.

The need to implement and derive sensible optimization technique to vary the manner the info is ruled. Manufactures are started synchronizing and analyzing the massive datasets for future call and to grow their business and production productively. Optimization technique has many capabilities that build it a perfect alternative for knowledge analysis in such state of affairs. Firstly, this method is intended for analyzing and drawing insights for extremely advanced system with immense knowledge volumes, multiples constraints and factors to be proclaimed for. Secondly, market has totally different

[1, 2]*Amity Institute of Information Technology, Amity University Uttar Pradesh, India*
E-mail: [1]skkhatri@amity.edu and [2]dikshadeo@gmail.com

range of enterprise objective related to it like value reduction, demand fulfillment etc.

A. Alternative of algorithmic program:

One among the foremost necessary things is to settle on acceptable algorithm for optimizing massive knowledge.

B. Steps to method massive data:

1. Classic ETL process.
2. Knowledge discovery and Investigate analysis: interactive knowledge exploration
3. Massive knowledge refinery: Store, mixture and remodel multi structured knowledge to correct worth.
4. Share refined knowledge and runtime modal.
5. Retain runtime modals and historical knowledge for current refinement and analysis.
6. Metallic element (Business intelligence) and analytics: retain historical knowledge to unlock extra worth. Like dashboard and good image analytics.

2.0 LITERATURE REVIEW

Distributed information Mining in Peer-to-Peer Networks (P2P) [1] offers associate degree summary of distributed data-mining applications and algorithms for peer-to-peer surroundings. It describes both exact and approximate distributed data-mining algorithms that work in a decentralized manner. It illustrates these approaches for the matter of computing and observation clusters within the information residing at the completely different nodes of a peer-to-peer network. This paper focuses on associate degree rising branch of distributed information mining known as peer-to-peer information mining. It additionally offers a sample of actual and approximate P2P algorithms for bunch in such distributed environments.

Web Service-based approach for information mining in distributed environments [2] presents associate degree approach to develop a knowledge mining system in distributed environments. This paper presents a net service-based approach to solve these issues. The system is engineered victimization this approach offers a uniform presentation and storage mechanism, platform autonomous interface, associate degreed an dynamic protractile design. The planned approach during this paper permits users to classify new incoming information by choosing one amongst the antecedently learnt models.

Architecture for data processing in distributed environments [3] describes system design for climbable and transportable distributed data processing applications. This approach presents a document image known as imp for accessing and looking out for digital documents in fashionable distributed info systems. The paper describes a corpus linguistic analysis of giant text corpora primarily based on collocations with the aim of extracting linguistics relations from unstructured text.

Distributed information Mining of giant Classifier Ensembles [4] presents a new classifier combination strategy that scales up expeditiously and achieves each high prognostic accuracy and trait of issues with high ramification. It accelerates a world model by discovering from the averages of the native classifiers output. The effective combination of enormous variety of classifiers is achieved this fashion. Multi Agent-Based Distributed information Mining [5] is the integration of multi-agent system and distributed data processing (MADM), additionally referred to as multi-agent primarily based distributed data processing. The angle here is in terms of implication, system's read, existing systems, and analysis tendencies. This paper presents an outline of MADM systems that are conspicuously in use. It additionally defines the common elements between systems and offers an outline of their methods and design.

Preserving Privacy and sharing the information in Distributed atmosphere victimization cryptographically Technique on rattled information [6] proposes a framework that enables systematic transformation of original information victimization randomized information perturbation technique. The changed information is then submitted to the system through cryptographically approach. This method is applicable in distributed environments wherever every information owner has his own information and needs to share this with the opposite information homeowners. At an equivalent time, this information owner needs to preserve the privacy of sensitive information within the records.

Distributed anonymous information perturbation technique for privacy-preserving information mining [7] discusses a light-weight anonymous information perturbation technique for economical privacy conserving in distributed data processing. 2 protocols are planned to deal with these constraints and to safeguard information statistics and also the organization method against collusion attacks. Associate degree algorithmic rule for Frequent Pattern Mining supported Apriori [8] proposes 3 completely different frequent pattern mining approaches (Record filter, Intersection and also the planned Algorithm) supported classical Apriori algorithmic rule. This paper performs a comparative study of all 3 approaches on a data-set of 2000 transactions. This paper surveys the list of existing association rule mining techniques and compares the algorithms with their changed approach.

Using Apriori-like algorithms for Spatio-Temporal Pattern Queries [9] presents a approach to construct Apriori-like algorithms for mining Spatio-temporal patterns. This paper addresses issues of the completely different sorts of examination functions that will be used to mine frequent patterns. Map-Reduce for Machine Learning on Multi core [10] discusses ways to develop a broadly applicable parallel programming paradigm that is applicable to different learning algorithms. By taking advantage of the summation kind during a map-reduce framework, this paper tries to pose a large vary of machine learning algorithms and reach a major speed-up on a twin processor cores. Victimization Spot Instances for Map cut back Work flows [11] describes new techniques to improve the runtime of Map cut back jobs. This paper presents Spot Instances (SI) as a means of attaining performance gains at low monetary cost.

3.0 PROPOSED ENHANCED APRIORI ALGORITHM

Apriori algorithm finds all frequent item sets by scanning the database time after time. This algorithm consumes a lot of time and memory space so to present parallelization in Apriori algorithm, an bettered Apriori algorithm is proposed which is shown below:-

1. In opening, parallel scan initial split the group action information horizontally into‘s’ node subsets and distribute it to‘t’ nodes supersets.
2. The various‘s’ nodes are then processed more. Main concern is to separate into major chunks of information that's reticulated inside.
3. Then once the method is completed, every node scans its own information sets then generates set of Candidate item set Kp.
4. Then the support count of every Candidate item set is about to one. This Candidate item set Kp is split into r partitions and sent to ‘r’ nodes with their support variety count. ’r’ nodes severally conglomerate the support variety count of an equivalent item set to provide the ultimate sensible support, and influence the frequent item set Fp within the partition once examination with the minimum support min_sup.
5. Once the process of every candidate item set, Kp is split more to calculate the frequent itemset exploitation pruning.
6. Finally merge the output of ‘r’ nodes to come up with set of world frequent item set K. This impermanent algorithmic rule is employed to significantly scale back the time as during this algorithmic rule obtaining frequent item sets by traversing the transactional information only once. The performance degradation with Sector filing system is perhaps thanks to I/O overhead wherever there's an excellent input and output transactions of file transfer between the native and distributed filing system. Another potential vital issue behind this could be JNI (Java Native Interface) overhead because it faces issue whereas researching JNI Layer to access Sector. Once merging we tend to not solely save time interval however additionally scales back the quantity of processors interactions throughout the execution of program.

There are some of the known methods to improve the efficiency of Apriori Algorithm using parallel dynamic Itemset count: Add new candidate item sets only when all of their subsets are estimated to be frequent. Otherwise, drop that item set or dump that data. [12]

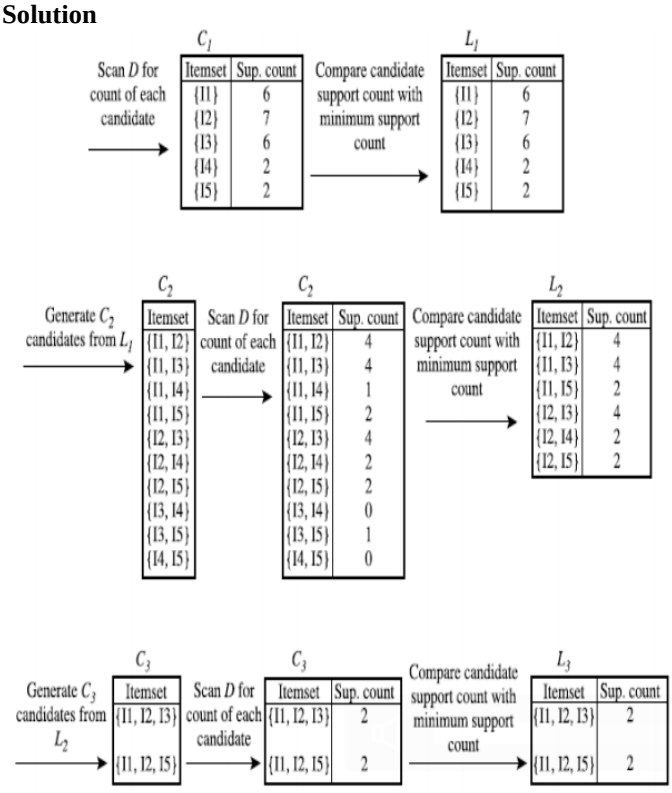
4.0 ANALYSIS OF EXISTING APRIORI ALGORITHM AND PROPOSED APRIORI ALGORITHM

In this part, we will compare time efficiency in finding frequent item set by normal Apriori algorithm and by method proposed in this paper. Now, consider the following example and calculate time to generate frequent item sets by using basic Apriori algorithm.

Problem: This problem justifies the legacy statement, how actually it works and process for a valid output. [10]

TID	List of item_IDs
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	
I1,I2,I3,I5	

Solution



As discussed, the existing Apriori algorithm has many loopholes, they are as follows:-

1. Historical Significance.
2. Suffers from a number of insufficiency and trade-offs.
3. Uses a minimum number of support thresholds.

According to the proposed methodology, the comparisons are based on using the native Java application. Table1. Indicates the results that will show how many records of n transaction can generate how many number of association rules.

Table 1: Proposed Apriori algorithm: No. of rules generation

No. of Records	Support	Confidence	Minimum support threshold	Generation time (seconds)	Number of frequent sets	No. of rules
1800	20	50	20% (360 records)	0.01	107	512
1750	20	50	20% (350 records)	0.01	107	425
51	20	50	20% (10.2 records)	0.00	55	210

```
Number of records = 1800
Number of columns = 10
---------------------------------------
INPUT SUPPORT THRESHOLD:
Support threshold (%) = 20.0
Minimum support (# records) = 360.0
---------------------------------------
INPUT CONFIDENCE THRESHOLDS:
Confidence threshold (%) = 50.0
---------------------------------------
SORT INPUT DATA:
---------------------------------------
APRIORI-GEN:
Minimum support threshold = 20.0% (360.0 records)
Generation time = 0.01 seconds (0.0 mins)
Number of frequent sets = 107.
---------------------------------------
```

After tested on various performance analysis portals, we came to a conclusion that the existing algorithm is not as efficient and reliable enough to use. To use it for big data optimization, we have to make sure that all parameters are covered properly and then we can implement further.

Table 2: Performance evaluation between existing and proposed algorithm

Itemset	Efficiency/ Performance Improvement	Existing Apriori Algorithm	Proposed Apriori Algorithm
800	20%	0.08 s	0.02s
1200	34%	0.20s	0.08s
1800	42%	0.40s	0.10s
2000	48%	0.50s	0.30s

5.0 IMPLEMENTATION OF PROPOSED METHOD

5.1 Pseudocode : Apriori Algorithm

The aprioriGen() accepts the set of all (k-1) large item set as parameters and returns a superset of the k large itemset as the candidate set. The outermost for Loop keep repeating k to further generate the candidate set for all levels of large itemset. Observing that the (k-1) itemset statement is detected as global huge set, thus we are not able to make a meaningful disseminated form of this function, and for sure, we can replicate the (k-1) large item sets among all the sites, but this is little obvious and thus makes no significance.

Step 1: Joining- Fk is generated by joining P(k-1) with itself.

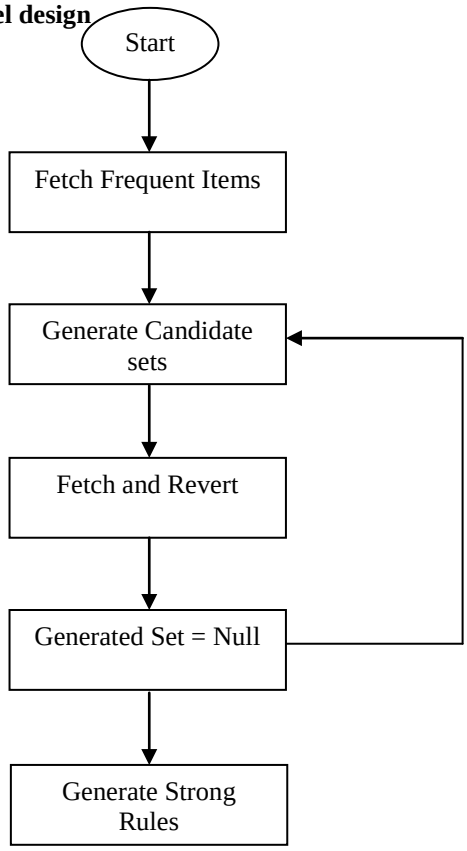
Step 2: Pruning- Any (k-1) - itemset that is not frequent cannot be a subset of a frequent k-itemset.

Step3: Generating Rules- frequent itemset then generates strong association rules.

Algorithm:

1. *pk: Candidate thing set P of size k.*
2. *sk: Frequent thing set S of size k.*
3. *f1 = {processed continuous items};*
4. *for(k= 1;sk!=Ø; k++) do Begin*
5. *ck+1= hopefuls created from Fk;*
6. *for every transaction t in database does augment the include of all applicants Ck+1 that are held in transaction t.*
7. *then, Fk+1 = applicants in Ck+1 with min_support limit esteem.*
8. *end;*
9. *return F*

5.2 High level design



6.0 IMPLEMENTATION WITH MAP REDUCE

The Map Reduce is a distributed Programming framework projected for massive cluster of systems arrangements that will operate in parallel on a really Brobdingnagian dataset. The task huntsman is to blame for managing the Map Reduce method. The tasks separated by the most application are foremost processed by the map tasks in a wholly parallel manner. The Map cut back framework kinds the maps output, that are then use as associate degree input for reducing the tasks. Each output and input of the roles are keep within the filing system. owing to parallel computing nature of Map cut back, parallelizing information mining algorithms exploitation the Map Reduce model has received important attention from the analysis community since the introduction of the model by Google. The Map cut back model supported Hadoop is examined for pertinence within the field of knowledge Mining. Steps to implement Enhanced proposed method with Map Reduce framework:-

Step 1: Maps the input dataset to N partitions, where N = number of slave machines.

Step 2: Reduce phase would take the immediate key-value pairs emitted in the map phase.

Step 3: Send them altogether to the master node for further collecting the number for count per item.

3.1 First, to generate frequent itemset in form of <item, count> key-value pair, which tells the number of occurrences.

3.2 Second, to generate the candidate sets from the source data file. It first prune those items that occur minimum than the support threshold by looking up the global Hash map list and then recursively call GenerateList() function.

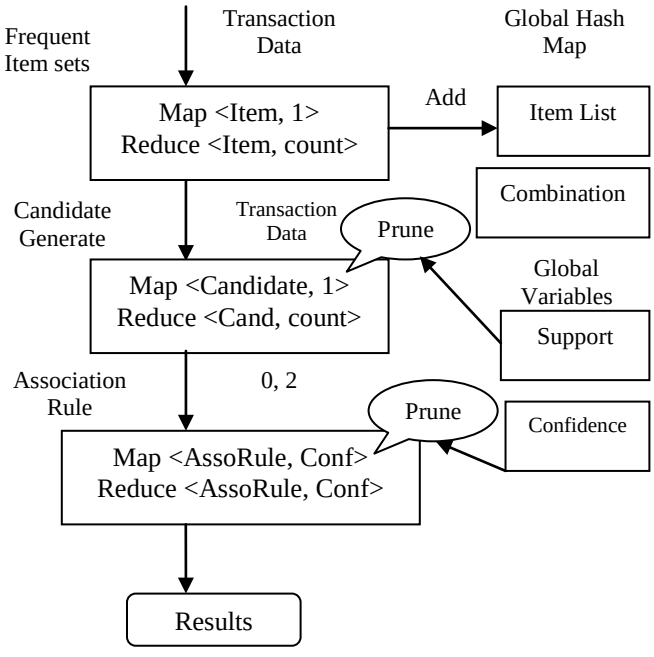


Figure 2: Block diagram to show working of Global Hash in Map Reduce.

7.0 CONCLUSIONS

Association Rule primarily based parallel information mining rule that deals with Hadoop Map Reduce. With this speedy detonation of information, process is preceded from terabytes era to pebibytes era. This trend produces the demand for progression advancement in data collection and storing technology. Thus there's a growing demand to run data processing rule on terribly giant datasets. Hadoop is that the computer code model framework for writing sensible applications that quickly method large amounts of information in parallel on immense clusters of computed nodes. It works on Map Reduce programming model. Map cut back could be a generic execution engine that parallelizes computation over an outsized cluster of machines.

8.0 FUTURE WORK

While experimenting with huge clusters using Hadoop, I came to a problem assertion that it will not disseminate any global variables to be apportioned by every partition due to its nature Sharing-nothing architecture. It can be implemented with Map reduce framework to provide more flexible development environment.

9.0 ACKNOWLEDGMENT

Authors express their deep sense of gratitude to the Founder president of Amity Universe, Dr. Ashok K. Chauhan for his keen interest in promoting research in the Amity University and has always been an inspiration for achieving great heights.

REFERENCES

[1]. Souptik Datta, Kanishka Bhaduri, Chris Giannella, Ran Wolff, and Hillol Kargupta, Distributed Data Mining in Peer-to-Peer Networks, University of Maryland, Baltimore County, Baltimore, MD, USA, Journal Ieeeinternet Computing chronicle Volume 10 Issue 4, Pages 18 - 26, July 2006.

[2]. ning Chen, Nuno C. Marques, and Narasimha Bolloju, A Web Service based methodology for information mining in dispersed situations, Department of Information Systems, City University of Hong Kong, 2005.

[3]. mafruz Zaman Ashrafi, David Taniar, and Kate A. Smith, A Data Mining Architecture for Distributed Environments, pages 27-34, Springer-Verlaglondon, UK, 2007.

[4]. grigoriostsoumakas and Ioannisvlahavas, Distributed Data Mining oflarge Classifier Ensembles, SETN-2008, Thessaloniki, Greece, Proceedings,companion Volume, pp. 249-256, 11-12 April 2008.

[5]. vudasreenivasarao, Multi Agent-Based Distributed Data Mining: Anover View, International Journal of Reviews in figuring, pages 83-92,2009.

[6]. p.kamakshi, A.vinayababu, Preserving Privacy and Sharing the Datain Distributed Environment utilizing Cryptographic Technique on Perturbeddata, Journal

Of Computing, Volume 2, Issue 4, ISSN 21519617, April2010.

[7]. feng LI, Jin MA, Jian-hua LI, Distributed unnamed information perturbationmethod for protection safeguarding information mining, Journal of Zhejiang Universityscience An ISSN 1862-1775, pages 952-963, 2008.

[8]. goswami D.n. et. al., An Algorithm for Frequent Pattern Mining Based On Apriori (IJCSE) International Journal on Computer Science andengineering Vol. 02, No. 04, 942-947, 2010.

[9]. marcingorawski and Paweljureczek, Using Apriori-like Algorithms for Spatio-Temporal Pattern Queries, Silesian University of Technology,institute of Computer Science, Akademicka 16, Poland, 2010.

[10]. cheng-Tao Chu et. al., Map-Reduce for Machine Learning on Multicore,cs Department, Stanford University, Stanford, CA, 2006.

[11]. navrajchohanet. al., See Spot Run: Using Spot Instances for Map-Reduce Workflows, Computer Science Department, University of California,2005.

[12]. Data Mining Textbook: Jawai Hen, and Michael Kamber.

An Alternative Approach in Generation and Possession of Backup Codes in Multi-Factor Authentication Scheme

Darren Pradeep D'Mello[1]

Submitted in June 2014; Accepted in March, 2015

Abstract - The paper describes the need for modification in the methods of generation and possession of backup codes in multifactor authentication schemes. The proposed system eliminates the need for storing/printing the printable backup codes. Here one-time verification code is displayed at the time of authentication. Which, the user has to modify by placing the pseudo key- digit from the pool of digits, at the pseudo key-position rendering authentication. This technique eliminates the risk of exhaustion of backup codes when all codes that were previously generated are used. Backup codes are only valuable to someone who had stolen the password.

Index Terms – Backup-codes, Multi factor authentication, OTVC, Pseudo Key-digit, Pseudo Key-position

1. 0 INTRODUCTION

Multi-factor authentication, MFA, There are several approaches in authentication schemes, M-FA is such a one which requires two or more of the three authentication factors [1]: a *knowledge* factor (KF - "something the person *knows*"), a *possession* factor (PF - "something the person *has*"), and an *inherence* factor (IF - "something the person is").

$$MFA = (KF \wedge PF) \vee (PF \wedge IF) \vee (KF \wedge IF) \quad (1)$$

Two-factor authentication 2FA, T-FA (or multi-factor authentication) is often misunderstood with "strong authentication". However, both are fundamentally different processes. Composite solutions from two or more of the three categories of factors will result into a True multifactor authentication [2].

KFs are the most common form of authentication widely used. In this form, the user or a person proves the knowledge of a secret to authenticate him. Secret in KF involves password or passphrase or PIN or pattern. In PFs security of the system relies on the physical protection of the PF itself and integrity of the authenticator, for example Mobiles, One-time pads, USB tokens etc. IFs determine the user is with the help of biometrics like fingerprint, voice, iris, face, DNA etc.

Proponents say that, In a T-FA, the incidence of online identity theft, and other online fraud, could drastically reduce because the victim's password would no longer be enough to access to the victims info by the hackers. [3]

In 2FA and MFA, PF is verified by One-Time Verification Code (OTVC). OTVC can also be referred as OTP based on the context of how the application being developed. In this paper, OTPs and OTVCs are used interchangeably. Text messaging (SMS) is a common technology used for delivering One-Time Password (OTPs) to the user. Since SMS is a universal correspondence station, being straightforwardly accessible in about every versatile handset, to any portable or landline phone, content informing has an incredible potential to achieve all buyers at a low aggregate expense. However, the cost of SMS for every OTVC may be unbearable to some users. In spite of threats from hackers, the mobile phone operator also contribute to form as a part of the trust chain[4]. Moreover, several mobile phone operators has to be trusted if a user is in roaming network, which may prone to mount a 'man-in-the-middle' attack.

As an illustration, as of late Google has begun offering check codes to versatile and landline telephones for all Google accounts. The user receives the OTVC either as a SMS or as an automated phone call using text-to-speech conversion. In case, if the users registered phone(s) are inaccessible, then the user can even use one of a set of (up to 10) previously generated one-time backup code (BC) as a second authorization factor instead of dynamically generated OTVC, after signing in [5] with their account password. The BCs are advised to be printed and held in a wallet which too imposes a security threat, when a hacker comes into possession.

2.0 EXISTING SYSTEM

OTVC Codes are uniquely crafted for an account when the users need them. Typically, OTVC generation algorithms use pseudo-randomness or randomness; they vary greatly in their generation and methods of delivering them. Codes are sent to user via text message, proprietary tokens etc. as discussed in the introduction. The system functions well until phone service is available. What if a person is travelling in a flight? Today, major email/web service providers allow users to print a set of BCs at the time of registration, to authenticate[6] them when phone is not available or lost. What if all the codes are exhausted? The user has to rely on "Authenticator" apps. There could be several reasons for its unavailability.

[1]*Department of Statistics & Computer Science, KVAFSU/College of Fisheries, Mangalore, Karnataka 575 002, India*
E-mail:darren@cofmangalore.org

3.0 PROPOSED SYSTEM

The proposed system eliminates the need for printing BCs while phone is out of reach. Unlike the existing system, set of BCs are not generated prior and given to the user for their possession, instead OTVC is displayed at the time of authentication. The user has to modify the OTVC code by placing a Pseudo Key-Digit (PK_d) from the pool of digits (L) at Pseudo Key-Position (PK_p) that are displayed during authentication.

The PK_d is derived from the PF; PK_p is derived from KF; both at the time of registration (when phone was available) ensuring a part of MFA scheme. This approach uses KF and derived PF at the time of authentication, PF is derived when phone is available.

4.0 SYSTEM IMPLEMENTATION

The system implementation requires modification in BC generation at the server and its transfer mechanism. Since scalability is a major concern, the entire process does not alter the mode of delivering OTPs when phone is available. Implementation changes are required only when phone is not available with the user.

Conventional OTVPs are generated using Time-based One-time Password Algorithm or Mathematical algorithms [7] and sent to the user via voice/SMS when phone is available, soon after the user's credentials (username and password) are verified. Here, this new approach requires no modification. Its effective use involves the following stages, which are illustrated below,

4.1 Registration Phase

Occurs when phone is available: The Single Sign-On (SSO) [8] credentials are accepted, the user is then presented (as an option or mandatory) of MFA scheme. If the system mandates, then he has to register his mobile number and provide recovery options available under the web service. The server verifies the mobile number immediately by sending an Initial Verification Code (IVC),

The user then enters IVC into registration page to verify the possession of his phone.

PK_p & PK_d registration: Now the user is requested to choose a PK_p, which he has to pick from the position values (L_p).

$$L_p = \{ k \mid 1 \le k \le f(x) \} \quad (2)$$

$$f(x)=\text{DIGITS(OTVC)}+1 \quad (3)$$

The DIGITS function in Eq. (3) returns the number of digits, its value is implementation dependent. OTVC composed of digits from Time-synchronized or Mathematical algorithm. Some web service providers use 8 digits as a standard size for a backup code. To avoid key logging, implementations using AJAX with drag drop functionality is preferred.

Now the server sends a single digit number to the user's verified mobile. The single digit number is a PK_d, Possible values are 0,1,2…..9.

$$L_d = \{ 0,1,2,3,4,5,6,7,8,9\} \quad (4)$$

Which the user has to place it into PK_p to get him authenticated. This step, verifies the user's PF.

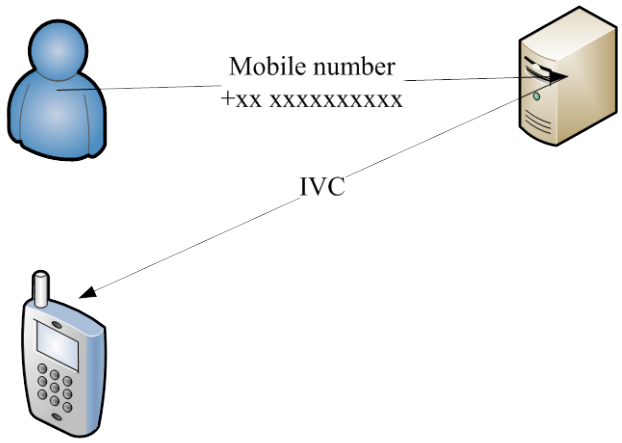


Figure 1: Registration phase

On verification, the user is taken to next step of registration.

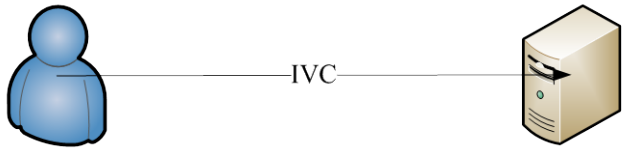


Figure 2: IVC confirmation

4.2 Utilization Phase

When phone is not available: User signs in by providing SSO credentials, and the user is directed to 2-step verification. Since the user has no phone service, he chooses alternate method of verification. Now the utilization phase is exercised as below. The user must now use the knowledge of PK_p and PK_d to complete the sign in.

Consider the example; the user gets OTVC in webpage after SSO as 46889513. This is a one-time code and by no means can the user expect to get the same at next sign in. A set of digits L_d is also presented as individual images[9] appearing as connected. Now the user has to place the PK_d at PK_p. The PK_p=k=4 and PK_d=7. Let us assume that the user's response will be PK_{ud} at position PK_{up}.

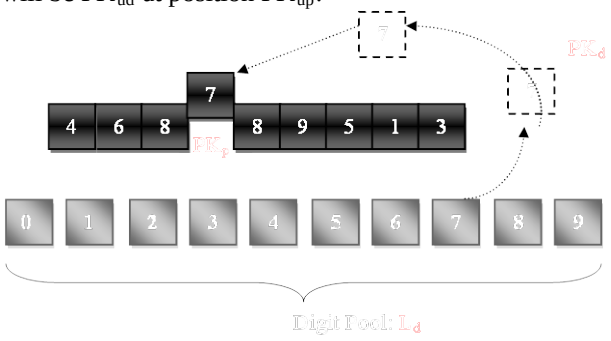


Figure 3: Process of placing PK_d at position PK_p by the user

On supplying PK_{ud} and PK_{up} the response is submitted, the integrity of the key-digit and key-position are verified to that of the user's registration and he is authenticated (if $PK_{ud} = PK_d$ and $PK_{up} = PK_p$). The utilization phase too is implemented using AJAX with drag drop functionality. PK_d and PK_p are stored at server at the time of registration.

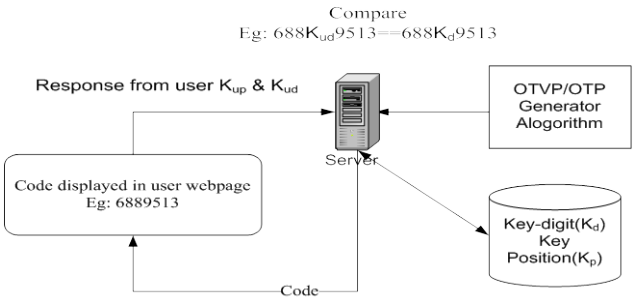


Figure 4: Block diagram of verification process

Any mismatch in key values i.e. if $PK_{ud} \neq PK_d$ or $PK_{up} \neq PK_p$ must prevent the user from signing in further. An invalid attempt limit may also be imposed. The user then has to recover using alternate options as configured by the Web service provider. At the server level, this can be summarized as show in fig 4.

The user must also be presented a choice for changing PK_d and PK_p soon after using them each time, when phone or voice service are available, however the aspiration is to reduce this repetitive overload.

5.0 FEASIBILITY EVALUATION

- It can be observed that, the existing overhead of generating a set of printable BCs at once and storing those in database at the server as well as printing by the user is not required. The cost is reduced.
- The later technique eliminates security risk associated at the user level of losing a printed sheet of BCs.
- Exhaustion of BCs never exists.
- The permutation and combinations of getting the right digit at right position is extremely high, Risk factor is minimum. CAPTCHA [10] may be used to prevent bots.
- Remembering PK_d and PK_p by the user is more effective than possession of a set backup codes.
- Overhead of downloading drag drop AJAX image (digit) is more than text, so workaround is required in minimizing the load.
- One-time passwords are vulnerable to social engineering attacks in which phishers steal OTPs by tricking customers into providing one or more OTPs that they used in the past [11].

6.0 FUTURE ENHANCEMENTS

Increasing the OTVC size or using an alternate alphanumeric algorithm will also enhance the technique.

7.0 CONCLUSION

In this paper, an alternative approach in generation and possession of backup codes in Multi-factor authentication scheme is discussed; although this mechanism cannot completely ensure the proper use of the system[12], it will surely reinforce the existing authentication scheme ensuring service availability to end user even while he is travelling eliminating the need of phone service or a set of backup codes.

REFERENCES

[1]. Wikipedia, Multi-factor authentication, 17 August 2013, http://en.wikipedia.org/wiki/2_factor_authentication

[2]. Federal Financial Institutions Examination Council, "Frequently Asked Questions on FFIEC Guidance on Authentication in an Internet Banking Environment", August 15, 2006

[3]. Two-Factor Authentication using Tivoli Access Manager WebSEAL, 06 Oct 2005, http://www.ibm.com/developerworks/tivoli/library/t-webseal

[4]. A. Chaudhary, V. N. Tiwari and A. Kumar, Analysis of Fuzzy Logic Based Intrusion Detection Systems in Mobile Ad Hoc Networks, *BVICAM's International Journal of Information Technology*, February 2014

[5]. Dilbag Singh and Ajit Singh, A Secure Private Key Encryption Technique for Data Security in Modern Cryptosystem, *BVICAM's International Journal of Information Technology*, December 2010

[6]. Mohammad Ubaidullah Bokhari and Shams Tabrez Siddiqui, A Comparative Study of Software Requirements Tools for Secure Software Development, *BVICAM's International Journal of Information Technology*, December 2010

[7]. RFC 4226, HOTP: An HMAC-Based One-Time Password Algorithm, http://tools.ietf.org/html/rfc4226

[8]. Rui Wang, Shuo Chen, and XiaoFeng Wang. "Signing Me onto Your Accounts through Facebook and Google: a Traffic-Guided Security Study of Commercially Deployed Single-Sign-On Web Services".

[9]. S.K.Muttoo and Sushil Kumar, Data Hiding in JPEG Images, *BVICAM's International Journal of Information Technology*, June 2009

[10]. Ahn, Luis von; Blum, Manuel; Hopper, Nicholas J.; Langford, John (2003). "CAPTCHA: Using Hard AI Problems for Security". *Advances in Cryptology — EUROCRYPT 2003*. Lecture Notes in Computer Science 2656. pp. 294–311. doi:10.1007/3-540-39200-9_18. ISBN 978-3-540-14039-9.

[11]. The Register article. The Register article (2005-10-12). Available: http://www.theregister.co.uk/2005/10/12/outlaw_phishing/

[12]. Neelabh, Tracking Digital Footprints of Scareware to Thwart Cyber Hypnotism through Cyber Vigilantism in Cyberspace, *BVICAM's International Journal of Information Technology*, December 2012

Exploring Sub Dominant Community on Web Graph: Using Link Structure and Usage Analysis

Nimisha Modi[1]

Submitted in June 2014; Accepted in March, 2015

Abstract - Information Retrieval (IR) process uses term based relevance measures to find relevant documents for a given query. Web IR utilities such as search engines tend to further process these relevant documents through link structure analysis and find rank score for each document within result set. The rank scores are used to sort the documents before presenting them to users for improving precision rate of top ranked results. Existing link analysis algorithms are using principal eigenvector of corresponding rank matrix for ranking. The limitation of using this approach is Web Local Aggregation (WLA). As an effect of WLA for the multi topic or polymorphic query, the dominant topic covers the major part within top ranked results and the sub-dominant topics are downgraded. The propose approach for link analysis serves for both ranking and grouping. Semantic analysis is incorporated along with link analysis to identify sub-dominant community through link analysis process and upgrade them according to the interest of the user searching on web. The paper suggests the model for search agent that categorizes the results based on their hyperlink connectivity and presents the groups of web pages that match semantic of user profile.

Index Terms – Eigenvector, Information Retrieval, Link Analysis, Usage profile, Web local aggregation, Web community

1.0 INTRODUCTION

Web contains huge amount of information. This information is available in form of millions of documents with varying degree of relevance. Specifically with respect to broad-topic queries, search engine returns a set of thousands or millions of documents as result. As the user's satisfaction requires few but highly authoritative results, the web search utilities encounter the challenge for retrieving only the most relevant resources in response to user query. This challenge raises the requirement to further process this result set before presenting it to user. The post-processing includes the expansion of result set, ranking of pages within result set, classification or clustering of result set etc... Basically, majority of the search engines assign the rank scores to web pages based on their hyperlink structure on web. As for example, the search engine Google pre-computes the PageRank [1] and presents the organized results to user in sorted order of PageRank scores.

Various research works are targeted towards statistical or machine learning algorithms for multi topic or polymorphic query. The broad topic query is a query for which various semantic are present on web documents. In response to broad topic query, normally the dominant topic or dominant semantic will appear among the top ranked results. Here, dominant topic or dominant semantic means the topic having more number of web documents that are highly connected with each other than that of documents with any other semantic.

For example, term 'java' is related with programming language java or island java. If we search web using query 'java', the index searching returns documents which contains term 'java' without concern of their semantic (either programming language or island or java coffee). These millions or billions of web documents returned through index searching of search engine are re-arranged by search engines in order of documents' ranking scores obtained through proprietary link analysis algorithm. Obviously, the major portion of result set present java in context of a programming language and these documents are even highly connected. As consequences, the documents with programming language 'java' cover the major portion in top ranked result set and user are not able to find those documents containing information about island 'java'.

My objective is to help users who search for diverse aspect of broad topic. I identify documents having various aspects of topic (i.e. various categories) with use of principal component analysis. Rather than using content of documents, I use their hyperlink structure for identifying groups of related documents.

2.0 RELATED RESEARCH

I analyze research works that is done in area of hyperlink structure analysis, web page categorization and semantic analysis based on user profile within information retrieval field. Hyperlink provides very rich information in addition to text that sometimes beats the text in form of quality and reliability. The role of hyperlink [2] is to confer the trust that one document puts on other document via the hyperlink to the later. Network of social interaction are found between web pages by hyperlink to other web pages. For exploitation and analysis of hyperlink structure on web graph [3], a number of algorithms have been introduced. All of them generally follow and possibly improve the concept of three basic algorithms: HITS [4], PageRank [1] and SALSA [5].

Basically these three algorithms work to assign a numerical weigh (rank) to each element within a hyperlinked set of documents on web graph.

These link analysis algorithms calculate rank based on varied type of neighborhood graphs that are represented by some adjacency matrix or transition probability matrix. They use the principal eigenvector [6] of matrix to assign the rank score for web pages. Google's PageRank is a variant of the eigenvector centrality measure.

[1]*Department of Computer Science, VNSGU, Surat, Gujarat.*
E-mail:nimishamehta@yahoo.com

Algorithms for web page categorization use different information containers such as - URL, unstructured text content, structure of the web page within markup tags, snippets i.e. short description of pages displayed with initial results of base search engine and linkage information in the form of incoming and outgoing links. Learning model for automatic classification [7, 8] is based on a combination of text and link analysis for distilling authoritative web resources.

Categorization of web pages is broad research area [9, 10, 11] where a huge variety of classification as well as clustering algorithms are introduced. As supervised learning or classification requires pre-specify topic taxonomy, the approach is limited for customized use only. Web directory are generally using such models to find the web page that match priory specified label. Majority of categorization algorithms are following unsupervised leaning i.e. clustering [10, 12, 13] to group web pages according to some intrinsic similarity within them either in term of content or structure.

Web usage mining [14] refers to the discovery of user access patterns from web usage logs. I analyze the research works which focus on use of usage profile before finding similarity of query phrase and documents. Hang and his collogues [15] proposed a method for query expansion by mining user logs, where the user profile is analyze prior to query is being submitted to the search utility and based on that the query is being expanded.

The limitation of such query expansion or refinement is that - the search becomes narrow (specific) search. Possibly user may require searching for some other perspective of topic. For example, a person in biology may require a search for type of virus that infected his laptop, in such case query expansion approach needlessly limits the search results to biological context and fails to satisfy user's requirement.

3.0 PROPOSED METHODOLOGY

I introduce some different approach that molds the application of link analyses algorithm to identify various community on web graph. For usage analysis, I prefer to utilize the user profile at post-link analysis stage rather than pre-processing stage as suggested in past studies [15].

3.1 Application of Link Analysis beyond Document Ranking

The limitation of ranking algorithms (PageRank [1] or HITS [4]) is that - they promote the highly interconnect dense set of nodes on web graph. As a consequence of web local aggregation (WLA) [16], the sub-dominant communities are downgraded by the dominant community. User could rarely find documents representing sub-dominant category in top 50 results and user are not even try to explore more than top 20-to-30 documents. My goal is to identify such sub-dominant communities and to upgrade them while presenting results based on interest of that user.

Existing link analysis algorithms are using only principal eigenvector of corresponding rank matrix. I am interested in higher order eigenvectors rather than only principal eigenvector. I take higher order eigen values and their corresponding eigenvectors. By looking at a larger set of eigenvectors, I find clusters of web pages that reflect through web local aggregation. The most important web community corresponds to the principal eigenvector and the component values within each eigenvector represent a ranking of web pages. The subsequent eigenvectors denotes corresponding minor communities on web graph

3.2 User Profile for Post-Query Processing

I suggest an approach of using users' profiles after the initial results are collected but before presenting these results to user. The approach is based on two objectives – first, we require broad search before any semantic analysis and then perform grouping of search results. The reason behind this approach is that we can present all the result as a broad query as well as specific to some semantic in different groups, as user's context of search query may be or may not be according to his previous search activity. With proposed approach, we use the user profile to make it convenient for user to find the topic of his routine at top results and to degrade the outliers that selected incorrectly within top results through link analysis process.

The second objective is to propose such a model that can be used by general purpose search engine. Evangelos traces [17] a popular search engine (Excite) to show the significant locality in the queries. More than 20%-30% of the queries have been previously submitted by the same or a different user. Search engines follow the practice to make a cache for query results for frequent queries. Research Experiments also shows the improvement in hit ratio via two level caching i.e. a cache of query result and a cache of inverted list of query terms [18]. So, in place of raw query result, results of link analysis can be stored as cache for frequent and broad topic query. When a search is performed on that query, it just needs to map the proper group of results. At time of user's search for such query, the semantic analysis is performed and results are collated from the cache copy.

3.3 System Model

Based on two approaches describe in section 3.1 and section 3.2, I propose the model for supporting system for web information seeker which now onwards will be referred as Search Assistant (SA). SA works as interface between user and search engine to monitor the search process and assist web users to intelligently retrieve information from the web. The model for the proposed system is described in figure 1.

This section outlines the flow of the system in three phases - collecting base result set, link analysis phase and semantic analysis phase.

- Collecting Base Result Set: User is provided the interface for interactive input for query that is submitted to search engines. It collects the initial results of user's query from search engines using the APIs from search engines like Google, Yahoo and MSN Search. Title and snippet returned by search engines is also stored along with URL.

This initial result set is referred as root set. The in-links and out-links of root result set are also collected. The collection of URLs in the root set, their in-links and out-links are collectively known as base set.

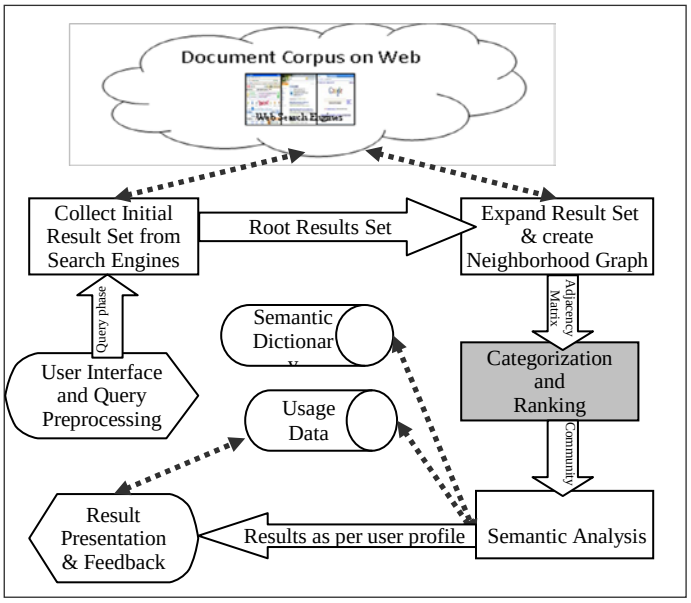


Figure 1: Proposed Model

- Link Analysis Phase: Link analysis algorithm visualizes this base set as the neighborhood graph N where each page (link) represents a node and a hyperlink from one page linking to another page is represented as directed edge. This neighborhood graph reflects linked structure of web pages in base set. We submit this neighborhood graph in form of adjacency matrix as an input for link analysis algorithm. Considering A as the adjacency matrix on neighborhood graph N, the authority matrix [6], AUT is derived from adjacency matrix A as $AUT = A^T * A$.
  SA applies principal component analysis (PCA) and make use of eigenvectors of adjacency matrix to identify the principal components i.e. web communities. So, next step is to calculate the set of eigen values and corresponding eigenvectors for authority matrix AUT.
  The eigenvector corresponding to highest eigen value is the principal eigenvector that denotes the dense set of web pages on web and thus interpreted as dominant community. The next higher order eigen value is corresponding to second dense set of web pages and thus second dominant web community and so on. We form the groups of web pages for each higher order eigenvector via collecting web pages having high rank score on corresponding eigenvector. Each group presents different web communities those pertaining to different semantics/topics.
- Semantic Analysis Phase: For personalization of results, SA filters the result set based on user profile. User profile basically describes the user's preference. The trivial approach to built user profile for IR is to create a set of words or word phrases that are frequently used by that user.

  SA collects the user profile using web access logs of users that are collected from web proxy server. In convention IR system, similarity measure is computed through matching documents to query phrase.

I suggest query post-processing technique ─ the set of documents that SA analyzes is already judged as query relevant by search engines. So to find similarity scores for documents, this IR model substitute query phrase with user profile. SA finds similarity of all groups with user profile and presents the set of web pages i.e. web communities according to the interest of user. Thus the search becomes general as well as specific search.

4.0 EXPERIMENTS AND FINDINGS

I select some broad topic or polymorphic terms and apply search using the user-id of users with different profiles. Form these experiments I pick 5 queries to discuss within this paper, which are polymorphic and leading to more than one semantic i.e. java, mouse, jaguar, virus, tree.
I identify dominant and sub-dominant set of results for selected topics. As per that, results of 'jaguar' on selected search engines categorized in to two categories: jaguar as well-known automobile manufacturing company as dominated category and jaguar as wild animal category. Similarly, categorization of results for query 'mouse' is animal mouse or an electronic input device. 'java' is categories as programming language and island. 'category of viruses' is classified as computer virus and biological virus. 'tree' is either woody plat or data structure .
I use the terminology RRS, LARS, SARS to evaluate the results of our experiments at different phases. RRS (Root Result Set) – precision rate for results returned by base search engines. LARS (Link Analysis Results Set) - precision rate for results set formed after link analysis and before semantic analysis. SARS (Semantic Analysis Results Set) - precision rate for results after semantic analysis.
The precision rates RRS, LARS and SARS of the results obtained for identified dominant and sub-dominant communities are given in table 1 and table 2.

Table 1: Search Precisions for dominant category

Query Term	RRS	LARS	SARS	LARS-RRS	SARS-RRS	SARS-LARS
java	90.00	86.00	98.00	-4.00	8.00	12.00
mouse	51.33	76.00	90.00	24.67	38.67	14.00
jaguar	52.00	78.00	98.00	26.00	46.00	20.00
category of virus	66.67	64.00	88.00	-2.67	21.33	24.00
tree	54.00	81.00	90.00	27.00	36.00	9.00
Average				14.20	30.00	15.80

The comparative analysis of table 1 and table 2 shows that for dominant community, link analysis phase shows average

14.20% increase in precision while semantic analysis boosts 30% increase in search precision than that of root results (RRS).

Similarly for sub-dominating community, link analysis shows average 42.27% increase in precision rate while semantic analysis boosts 58.87% increase in search precision than that of root results (RRS).

Table 2: Search Precisions for sub-dominant category

Query Term	RRS	LARS	SARS	LARS-RRS	SARS-RRS	SARS-LARS
Java	6.00	50.00	68.00	44.00	62.00	18.00
mouse	33.33	70.00	85.00	36.67	51.67	15.00
jaguar	22.67	66.00	90.00	43.33	67.33	24.00
category of virus	21.33	60.00	80.00	38.67	58.67	20.00
Tree	11.33	60.00	66.00	48.67	54.67	6.00
Average				42.27	58.87	16.60

Improvement from LARS to SARS is 16.60% in case of dominating community; while in case of subdominant community it is 15.80%. So the role of semantic analysis is stable in both the case. These findings illustrate the significance of link base categorization, as we conclude that the categorization of grouping using link analysis plays a major role to boost sub-dominating community on web.

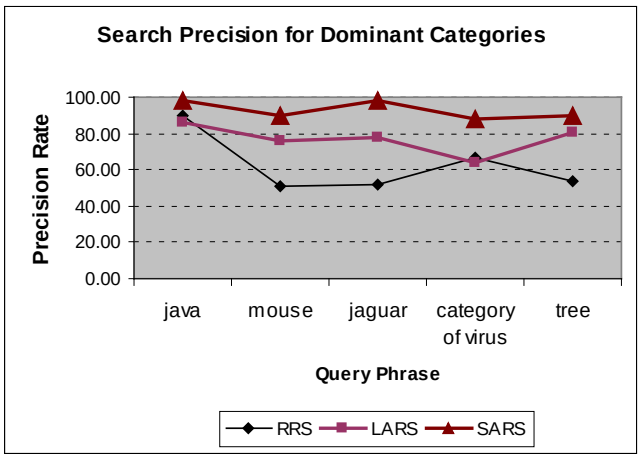


Chart 1: Search Precision for Dominant Category

Link analysis finds web communities based on hyperlink structure. It finds the sub-dominated communities. The role of semantic analysis is to identify the proper web community based on user's interest.

We can visualize the effect of both link analysis as well as semantic analysis on above results with chart 1 and chart 2.

Link analysis finds web communities based on hyperlink structure. It finds the sub-dominated communities. The role of semantic analysis is to identify the proper web community based on user's interest.

We can visualize the effect of both link analysis as well as semantic analysis on above results with chart 1 and chart 2.

5.0 CONCLUSIONS & FUTURE PATH

The proposed model is targeted towards broad topic query that applies PCA with eigenvectors for finding the group of web pages having common interest. Dominate category is having maximum resources than that of other. Dense linkage structure as well as rich text content of such dominating category suppressed the sub-dominating category via web local aggregation. SA uses the link structure of base result set to identify such set of highly linked group of web pages and using the principal component analysis we identify the dominated as well as some sub-dominated groups. Among these groups of documents, it performs the semantic analysis i.e. based on user profile.

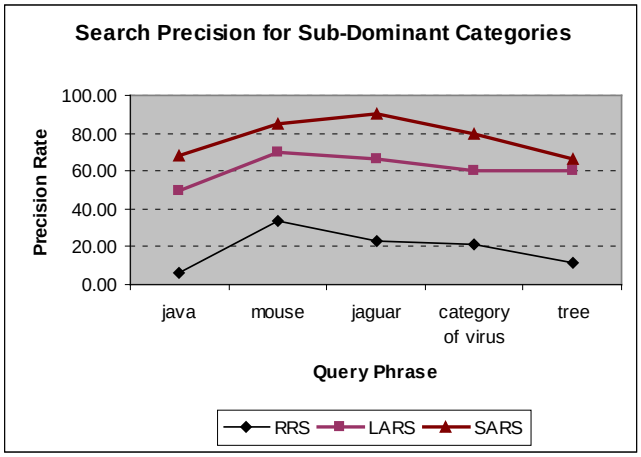


Chart 2: Search Precision for Sub-Dominant Category

The major limitation of algorithm is the central assumption that 'a hyperlink confers authority' which is largely applicable for social networks of the academic publications, but it is not guaranteed for commercial web pages. Sometimes, web sites are generally designed by commercial developers who link up their customers in densely connected cliques even though those customers have nothing in common

The algorithm presented in the paper generates the flat clustering on different themes. This can be enhanced towards hierarchical grouping in tree of topic [19] by analysis of inter-connection between documents on neighborhood graph.

Image, multimedia and other embedded objects are big sources of information, which are ignored except the text (e.g. anchor text, alternate text etc...) within their container markup tags. This indicates the other direction for future enhancement of the given model.

REFERENCES

[1]. Sergey Brin, Lawrence Page, "The anatomy of a large-scale hyper-textual Web search engine", In Computer Networks and ISDN Systems, 1998, 33:107-117.

[2]. Mandar R. Mutalikdesai, Srinath Srinivasa, "Co-citations as citation endorsements and co-links as link

endorsements", Journal of Information Science", v.36 n.3, p.383-400, June 2010

[3]. Lei Tang, Huan Liu, "Managing and Mining Graph Data, Advances in Database Systems", 2010, Volume 40, 487-513, DOI: 10.1007/978-1-4419-6045-0_16

[4]. Jon Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 1999, 46:604-632

[5]. R. Lempel, S. Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect", 9th International WWW Conference, 2000

[6]. Amy N. Langville, Carl D. Meyer, "The Use of Linear Algebra by Web Search Engines", Bulletin of the International Linear Algebra Society, 2005, 33:2-6.

[7]. Soumen Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, "Automatic resource list compilation by analyzing hyperlink structure and associated text", Proceeding of 7th International World Wide Web Conference, 1998.

[8]. Ajay S. Patil, B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", Proceedings of The International MultiConference of Engineers and Computer Scientists, IMECS 2012, March 14-16, 2012, Hong Kong, VOL I, Page No.

[9]. Xiaoguang Qi, Brian D. Davison, "Web Page Classification: Features and Algorithms", in Technical Report. 2007, Department of Computer Science and Engineering, Lehigh University: Bethlehem, PA. p. 1-31.

[10]. Gulli A., "On Two Web IR Boosting Tools: Clustering and Ranking", PhD Thesis, University of Pisa, May 2006

[11]. Ghanshyam Singh Thakur, Dr. R. C. Jain, "NFCKE: New Framework for Document Classification and Knowledge Extraction", BIJIT - BVICAM's International Journal of Information Technology, January-June, 2009, Vol.1 No.1; ISSN 0973-5658

[12]. Deepak P, Deepak Khemani, "Unsupervised Learning from URL Corpora", Proceedings of the 13th International Conference on Management of Data (COMAD-2006), Delhi, India

[13]. Anil Kumar Pandey, T. Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2009, Vol.1 No.2; ISSN 0973-5658

[14]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations. Vol. 1–2. 2000

[15]. Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma, "Query Expansion by Mining User Logs", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp. 829-839, July/Aug. 2003.

[16]. Ziyang Wang, "Improved link-based algorithms for ranking web pages", proceeding of 5th International Conference of Web Age Information Management, 2004.

[17]. Evangelos P. Markatos, "On Caching Search Engine Query Results", In Proceedings of the 5th International Web Caching and Content Delivery Workshop, May 2000.

[18]. Amarjeet Singh, Mohd. Hussain, Rakesh Ranjan, "Two Level Caching Techniques for Improving Result Ranking", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2011 Vol.3 No.2; ISSN 0973-5658

[19]. Parul Gupta, A.K. Sharma, "A Framework for Hierarchical Clustering Based Indexing in Search Engines", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2011 Vol.3 No.2; ISSN 0973-5658.

A Multimodal Approach to Improve the Performance of Biometric System

Chander Kant[1]

Submitted in January 2015; Accepted in April, 2015

Abstract - Biometric systems have very success rate in identifying an individual based on ones biological traits. In biometric history some features like weight, age, height etc. are also there to provide user recognition to some extent but not fully upto the mark because of their changing nature according to time and environment. These features are called soft biometric traits. Soft biometric traits are lack of permanence but they have some positive aspects in respect of enhancing the biometric system performance. Here in this paper, we have also highlighting the similar point but with a new aspect that is integrating the soft biometrics with fingerprint and face for improving the performance of biometric system. Here we have proposed an architecture of three different sensors to evaluate the system performance. The approach includes soft biometrics, fingerprint and face features, we have also proven the efficiency of proposed system regarding FAR (False Acceptance Ratio) and total response time, with the help of MUBI tool (Multimodal Biometrics Integration).

Index Term - primary biometric, soft biometric, FAR, minutiae point, multimodal biometrics.

1.0 INTRODUCTION

Biometric systems automatically recognize the individuals based on their physiological and behavioural characteristics such as hand-geometry, fingerprint, iris, face, retina, voice, palmprint, signature, gait pattern and keystroke dynamics [1]. Unimodal biometric system used to recognize only a single trait. There are a number of problems such as noisy sensor data, non-universality and lack of distinctiveness of the chosen biometric trait, unacceptable error rates, and spoof attacks. On the contrast, Multimodal biometric systems help in solving the problems associated with unimodal biometric systems. In multimodal systems evidence can be obtained from multiple sources [2]. A multimodal biometric is expected to be more reliable than unimodal system. Multimodal system will have two major limitations. The first one is very high quality sensors and very large database is required that increases overall cost and the second limitation is Verification time is increased which causing inconvenience to the users.

Due to the no. of limitations, the identifiers in a multimodal biometric system are generally restricted to two or three. One of the possible resolutions to this problem is to use soft biometrics like height, weight, age, and eye colour. The information obtained from the soft biometrics is indistinctive, not reliable, and can be easily spoofed so it is not enough to establish the identity of a person. This paper describes an approach for integrating the information provided by the soft biometric traits (weight/height) with the input of the primary biometric system. The performance increase obtained from this integration with the input of a fingerprint and face biometric system (multimodal) is analyzed.

If the characteristics of soft biometric can be automatically extracted and used during the decision making process then the overall performance of the system will improve and the need of manual involvement will be reduced. Rest of the paper is organised as follows: section II describes related work, section III shows the proposed scheme, section IV describes mathematical formulas, section V extracts soft biometric, comparison of proposed scheme with existing biometric technologies explains in section VI, section VII shows result and finally conclusion and future prospect in section VIII.

2.0 RELATED WORK

The purpose to design biometric system for the identification of criminals [3]. Basically, the identification was based on three sets of features. First is an Anthropometric measurement: height of the arm, second, Morphological description: appearance and body shape like eye colour and anomalies of the fingers, and third is Peculiar marks: moles and scars observed on the body. This system was useful in tracking criminals but as the features like weight, height, age, gender are common and temporary. So this system had an unacceptably high error rate and not acceptable. The soft biometric traits can be classified into two categories: first is continuous (height, weight) and another is discrete (age, gender, eye colour) [4]. It was shown that a combination of personal characteristics like age, gender, eye colour, height, and other visible identification marks can be used to identify an individual only by a limited accuracy [5]. The use of soft biometric traits like gender and age, for organize a huge biometric database was invented later [6]. As shown below Figure 1 explains the fusion of soft biometrics (like height/weight) such as weight with the primary biometrics (finger print, face). Filtering is the process of restrictive the number of entries in a database to be searched that can be based on characteristics of the interacting user. For example, if gender of the user can somehow be identified, the number of search entries will be improved as the search can be restricted only to the subjects with this profile enrolled in the database. This greatly improves the speed or the search efficiency of the biometric system. Filtering and system parameters tuning require an accurate classification of a user into a particular class [7]. The accuracy and performance of multimodal biometric authentication systems is examined using

[1]*Assistant Professor, Department of Computer Science & Applications, K.U., Kurukshtra, Haryana, INDIA.*
E-mail:ckverma@rediffmail.com

state of the art Commercial Off-The-Shelf products [8]. Fusion of ear and soft- biometrics results in an improvement of approximately 5.59% over the primary biometric system i.e. ear [9]. Experiments on the MSU and NIST multimodal databases show that fusion rules achieve consistently high performance without adjusting for optimal weights for fusion and score normalization on a case-by-case basis [10]. A biometric approach can be used for continuous user authentication by fusing hard and soft traits [11]. The continuous user authentication using soft biometric traits can be used for E-learning purpose [12].

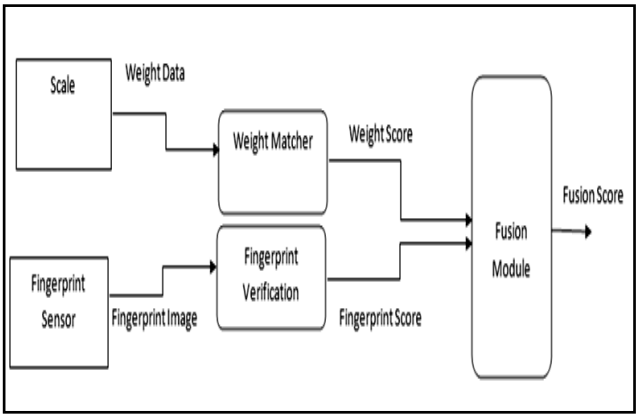


Figure 1: Fusion of Soft biometrics and primary biometrics

2.1 Soft Biometric Extraction

For using soft biometrics, a mechanism should be there to automatically (i.e. without user interaction) extract features from the user during the recognition phase [13]. This can be achieved using a particular system of sensors. For example, a collection of infrared beams could be used to measure the height, weighing machine can be used to measure the weight, a camera could be used for obtaining the facial image of the user, which can be used to obtain information like age, gender, and ethnicity [14]. The information obtained from soft biometrics could be used to count the identity information provided by the user's primary biometric identifier. Extensive studies have been made to identify the gender, ethnicity, and pose of the users from their facial images. The gender, ethnicity and pose of human faces are classified using a combination of experts by radial basis functions [15]. Their gender classifier can classify users as either male or female with a more than accuracy rate of 96 %. Age determination is a very difficult problem because physiological or behavioural changes in the human body are very limited as the person grows from one age to another [16]. Currently there are no reliable biometric indicators for age determination.

3.0 PROPOSED SCHEME

The proposed system combines the soft traits with fingerprint and face to get faster response time. Proposed method, as shown in figure 2, works by first comparing and matching soft biometric traits. If the result is not matched then it will directly rejects the user and if soft traits are matched then fingerprint and face traits are captured with the help of sensor. After that the feature sets of fingerprint and face are extracted. The system compares these values with the existing values in the database, and generates match scores of the respective traits. The system also generates the match score of soft trait. These match scores go through a (min-max) normalization process as explained in following section. After that the simple sum rule fusion technique is applied and a fusion match score is generated. If the resulting fusion score is equal to or above the set threshold value then the user is verified otherwise the system blocks the access and designate the user as imposter. There are number of advantages of proposed approach over the conventional system, as discussed below:

(i) The feature set of fingerprint and face are being calculated if and only if the user is found to be genuine at first stage (i.e. soft biometric phase)
(ii) This system improves the FAR(false accept rate)
(iii) Performance of the system improves.

The parallel execution of the process results in improved false acceptance rate. Though the proposed system improves the overall system performance yet it's not free from some drawbacks.

(i) Extra storage space is needed to store the templates with soft trait data like age, gender, height.
(ii) Total response time of the system increases for genuine user.
(iii) As the soft trait varies over period of time, the system must be used within those particular time period for which it remains invariant.

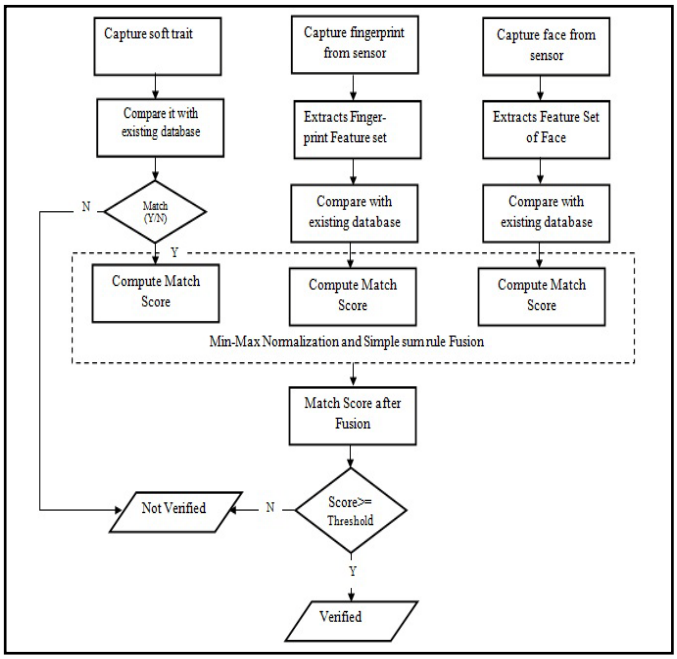


Figure 2: Proposed Scheme Architecture

Algorithm for verification/identification in proposed scheme

1) Capture Soft trait
2) Compare it with existing database
3) If (soft trait feature matched)
4) Compute match score of soft trait
5) Capture fingerprint from sensor
6) Extract fingerprint feature set
7) Compare with existing database
8) Compute fingerprint match score
9) Capture face from sensor
10) Extract feature set of face
11) Compare with existing database
12) Compute face match score
13) Apply (min-max) normalization on Soft, Finger and Face match scores
14) Apply simple sum rule fusion on normalized scores
15) If (fusion score >= threshold)
16) Verified/Identified
17) Else
18) Not Verified/Identified
19) End If
20) Else
21) Not Verified/Identified
22) End If
23) End

4.0 MATHEMATICAL FORMULAS

We are describing here in (min-max) normalization for score normalization and fusion formula to fuse the different modalities.

Let's matching score set is denoted as $\{S_k\}$ and normalized scores as $\{S'_k\}$:

- Min max normalization is top suited where the Upper and lower bounds (maximum and minimum values) of the scores produced by the matcher are known. This method is not vigorous; therefore, it is highly sensitive to outliers. [17]

$$S'_k = \frac{(S_k - \min)}{(\max - \min)}$$

If S_i is the matching score from i^{th} modality, S represents the resulting fused score.

- The *Simple Sum Rule* adds the scores as a linear transformation.

$$S = (a_1 S_1 - b_1) + \ldots + (a_n S_n - b_n)$$

a_i and b_i represents the weights and biases, respectively, which can be entered by the user.

5.0 COMPARISON OF PROPOSED SCHEME WITH EXISTING TECHNOLOGIES

The proposed scheme was implemented using MUBI software. Min-Max normalization and simple sum rule fusion method were used in the proposed scheme. The sample biometric data was taken from NIST website which is well recognized standards organisation. Figure 3-6 shows the comparison of proposed scheme on the basis of genuine acceptance rate and false acceptance rate.

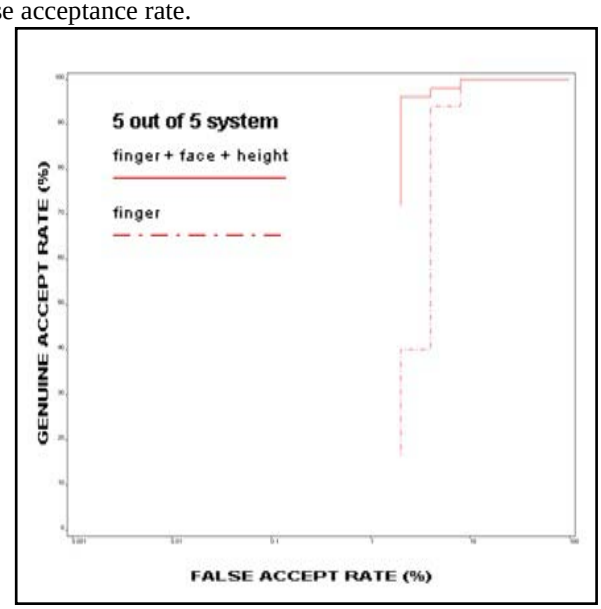


Figure 3: Proposed Scheme (i.e finger + face + height) v/s fingerprint system

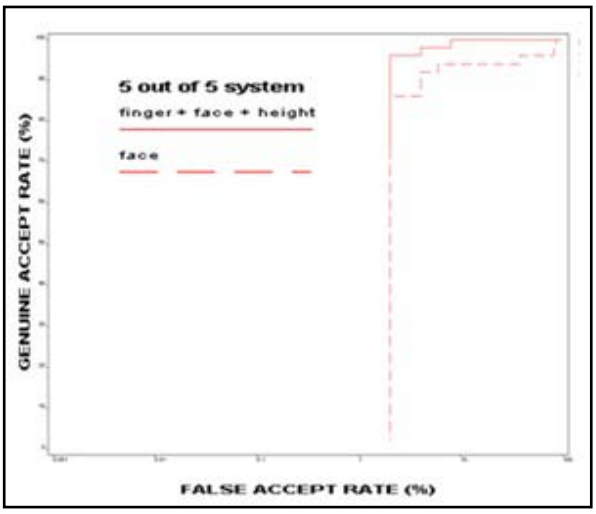


Figure 4: Proposed Scheme (i.e finger + face + height) v/s face system

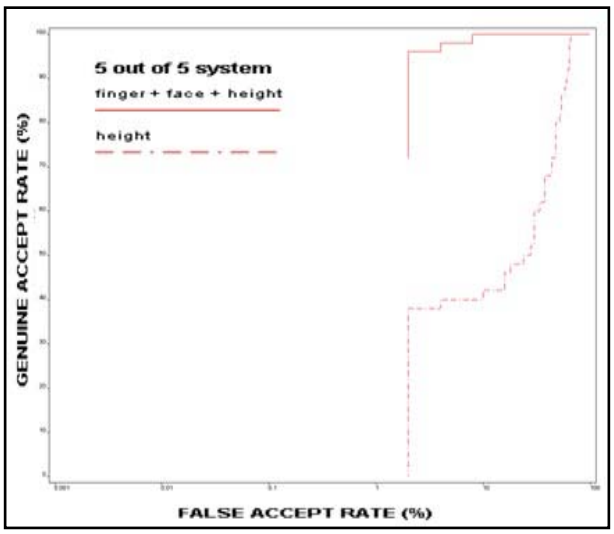


Figure 5: Proposed Scheme (i.e finger + face + height) v/s height system

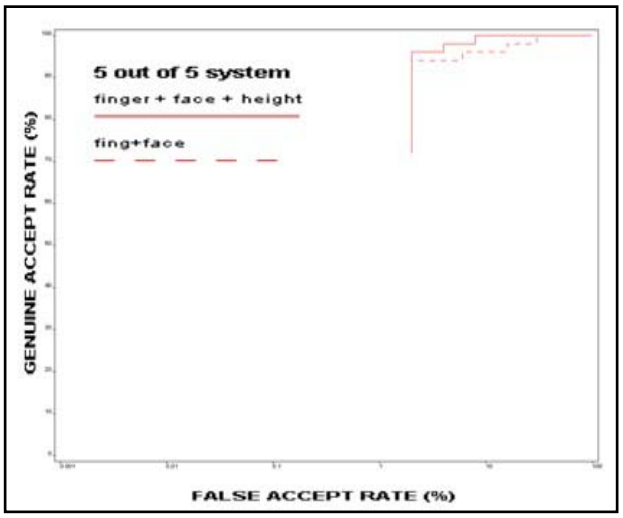


Figure 6: Proposed Scheme (i.e finger + face + height) v/s multimodal (i.e. finger + face) system

6.0 RESULT

Table 1: Comparison of proposed scheme with existing biometric techniques

S.No.	Biometric Technologies	GAR	FAR
1	Finger	72	3.946
	Proposed Scheme		1.859
2	Face	92	3.946
	Proposed Scheme		2.031
3	Height	72	47.106
	Proposed Scheme		2.051
4	Multimodal (finger+face)	95	5.623
	Proposed Scheme		1.943

It can be concluded from table 1 that the proposed scheme has improved false acceptance rate as compared to the other stated techniques.

7.0. CONCLUSION AND FUTURE SCOPE

This paper presents a simple and effective method of multimodal biometric authentication scheme based on soft biometric trait i.e. combination of fingerprint and face verification system. The proposed scheme shows that the soft biometric information such as blood group, gender, height, and age when combined with primary biometric traits will improve the performance of the traditional biometric systems. Methods to integrate time varying soft biometric information such as height and weight into the expected biometric framework is studied. This scheme allows us to completely control and automate fingerprint and face authentication with effective response time and FAR (False Accept Rate). Designed proposed scheme is not free from all loopholes. One of the negative aspects is that database will be very large due to accommodation of all the weight/ height, fingerprint and face template, therefore extra memory will be needed to store templates. As false rejection ratio is high in soft biometrics in future view, FRR can be improved and also a method can be developed for automatic extraction of soft biometric traits.

REFERENCES

[1]. B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan Studies of Biometric Fusion. Technical Report NISTIR 7346, NIST, September 2006.

[2]. A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, 14(1):4–20, January 2004.

[3]. Bertillon, A.: Signaletic Instructions including the theory and practice of Anthropometrical Identification, R.W. McClaughry Translation. The Werner Company (1896)

[4]. A. K. Jain, K. Nandkumar, X. Lu and U. Park, ―Imtegrating faces, fingerprints and soft biometrics traits for user recognition‖, in proceedings of ECCV international workshop on biometric authentication, volume LNCS 3087, pages 259- 269, prague, czeh republic, springer,may 2004.

[5]. Heckathorn, D.D., Broadhead, R.S., Sergeyev, B.: A Methodology for Reducing Respondent Duplication and Impersonation in Samples of Hidden Populations. In: Annual Meeting of the American Sociological Association, Toronto, Canada (1997)

[6]. Wayman, J.L.: Large-scale Civilian Biometric Systems – Issues and Feasibility. In: Proceedings of Card Tech / Secur Tech ID. (1997)

[7]. S. C. Dass, K. Nandakumar, and A. K. Jain. A Principled Approach to Score Level Fusion in Multimodal Biometric Systems. In Proceedings of Fifth International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), pages 1049–1058, Rye Brook, USA, July 2005.

[8]. K.Sasidhar, Vijaya L Kakulapati, Kolikipogu Ramakrishna and K.KailasaRao, Multimodal Biometric Systems – Study to improve accuracy and performance,

In: International Journal of Computer Science and Engineering Survey(IJCSES) Vol. 1, No. 2, November 2010

[9]. Shrikant Tiwari, Aruni Singh and Sanjay Kumar Singh, Fusion of Ear and Soft-biometrics for Recognition of Newborn, In: Signal and Image Processing : An International Journal (SIPU) Vol. 3, No. 3, June 2012

[10]. Sarat C. Dass, Karthik Nandakumar and Anil K. Jain, A Principled Approach to Score Level Fusion in Multimodal Biometric Systems, In: proceedings of AVBPA 2005

[11]. A. Prakash, A Biometric Approach for Continuous User Authentication by Fusing Hard and Soft Traits, In: International journal of Network Security, Vol. 16, No. 1, PP. 65-70, Jan. 2014

[12]. Kalyani Tukaram Bhandwalkar and P. S. Hanwate, Continuous User Authentication Using Soft Biometric Traits for E-Learning, In : International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Special Issue 4, April 2014

[13]. X. Chen, P. J. Flynn, and K. W. Bowyer. IR and Visible Light Face Recognition. Computer Vision and Image Understanding, 99(3):332–358, September 2005.

[14]. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition? In: Proceedings of SPIE International Symposium on Defense and Security: Biometric Technology for Human Identification . (2004)

[15]. E. Erzin, Y. Yemez, and A. M. Tekalp. Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability. IEEE Transactions on Multimedia, 7(5):840–852, October 2005.

[16]. Jain, A.K., Dass, S.C., Nandakumar, K.: Integrating Faces, Fingerprints, and Soft Biometric Traits for User Recognition. Proceedings of Biometric Authentication Workshop, LNCS 3087, pp. 259-269, Prague, May 2004.

[17]. A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition”, IEEE Transactions On Circuits And Systems For Video Technology, vol. 14, no. 1, pp. 4–21, January 2004

Predicting for Sustainable Insurance with Adaptive Gradient Methods

Parveen Sehgal[1], Sangeeta Gupta[2]and **Dharminder Kumar[3]**

Submitted in January 2015; Accepted in March 2015

Abstract - This paper extends the comparison of gradient based training methods used in the construction of prediction models based upon neural network, for sustainable insurance. Here adaptive gradient based techniques are compared with simple first order gradient based technique and with some second order training techniques for learning of the network. Convergence towards minimum error, for a number of first and second order algorithms are compared while utilizing data taken from live data warehouse of life insurance. Method of back propagation of errors is adopted for training of multilayer feed forward networks, while employing these gradient based algorithms of training error reduction. This paper is extended version of the paper presented in IEEE Conference INDIACom-2014.

Index Terms – Adaptive gradient algorithms, Error back-propagation, Error gradient, Multilayer-perceptron, Neural network, Supervised training, Sustainable insurance.

NOMENCLATURE

SDI – Sustainable Development Indicator
GD – Gradient Descent
CGM – Conjugate Gradient Method
SCGM – Scaled Conjugate Gradient Method
LMA– Levenberg Marquardt Algorithm
GDA – Gradient Descent with Adaptive Learning Rate
GDM – Gradient Descent with Adaptive Momentum

1.0 INTRODUCTION

Insurance industry need to exploit new market opportunities that will come from the shift of economic development to sustainable development. It has the power to inspire and create a greener society and plays a critical role to improve the adverse economic, social and environmental consequences of financial losses arising from fortuitous and accidental events. Sustainable Insurance provides a solution to inflate the innovative risk management and insurance ways that we need to promote saving of natural resources and to shape a safe future for the coming generations. One of the major goals is to offer low premium insurance policies and to support the rural communities in developing countries and to offer retirement planning solutions across the globe especially in rural areas.

Insurance in rural areas is a key sustainable development indicator (SDI) to safeguard the future of people in rural areas, which will otherwise have adverse effects on the usage of Based on the historical data; we have an urgent need to predict the prospective customers in rural areas and to launch new policies to ensure sustainable development. But due to complexity of nonlinear relationships present between input and output variables in the bulk amount of historical data and these types of problems are difficult to solve with older techniques. Various technologies of soft computing like Fuzzy Logic, genetic algorithms, evolutionary algorithms and artificial neural networks can be used to create the prediction models; which is related to field of nonlinear optimization [1][5]. The idea is to find an optimal solution for the complex nonlinear relationship between input and output variable; which generally exists in high dimensional data of real life problems [8].

In this paper, we extend the construction of prediction models based upon neural network and trained with adaptive gradient based techniques to find an approximation for the complex nonlinear relationship present in the insurance data, within the desired limits of accuracy [15]. Neural Networks are employed to solve complex problems of optimization in the areas of software engineering and data mining [17]. The gradient based algorithms used for the training of network optimize the weights present between the layers of the network until a state of minima of error gradient is reached. We descend along a multi-dimensional error gradient surface in a direction towards the minima of gradient, and proceed iteratively in small steps until we reach at the point of minima of the error gradient. We extend the study of gradient based training and compare the convergence and accuracy of adaptive gradient based techniques with first order and second order techniques.

2.0 LITERATURE SURVEY & LEARNING METHODOLOGIES

Taylor approximation of the error energy present during neural network training is written as:

$$E(W_k + z) \approx E(W_k) + E'(W_k)^T z + \tfrac{1}{2} z^T E''(W_k) z$$

When higher order terms are neglected then, first order and second order approximations for E is written as:

$$E_1(W_k + z) \approx E(W_k) + E'(W_k)^T z$$

$$E_2(W_k + z) \approx E_1 + \tfrac{1}{2} z^T E''(W_k) z$$

For minimizing of error function based on the weight vector present in neural network, we have used a number of first and second order approximation techniques.

[1]*Research Scholar, Dept.of CSE, NIMS University, Jaipur, India, (Corresponding Author), Tel:+91-9896931768.*
[2]*Prannath Parnami Institute for Professional Studies, Hisar, India.*
[3]*Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India.*
E-mail: [1]parveensehgal@gmail.com,
[2]sangeet_gju@yahoo.co.in and [3]dr_dk_kumar_02@yahoo.com

Steepest descent (simple gradient descent) uses first order information to descend toward the point of local minimum. But this method shows a poor convergence and uses fixed value of learning rate parameter and is not useful for practical applications that require a faster speed for output [2]. A careful selection of the step size is extremely important for faster convergence toward point of minima. With larger values of the step size it will diverge and if taken too small it will take longer times to converge. Also, complex shape of the multi-dimensional error surface usually presents irregularities and makes moving toward global minima problematic. Researchers have suggested modifications to improve the convergence and avoid stagnant learning and oscillations by introducing methods with varying step size. Proposed methods utilize learning rate and momentum parameters to decide for the optimal step size [18, 19].

When error is computed according to second order approximation then first order information is updated and a second order minimization step is performed. We have also applied second order methods like conjugate gradient, scaled conjugate gradient, which have their own positives and negatives in terms of computer memory required, convergence speed and accuracy.

Earlier methods like Newton's method shows a faster convergence toward point of minimum training error than first order methods, because it also utilizes second order information and approximates the second order terms by evaluating the hessian matrix [4]. But computation of hessian becomes very tedious and time consuming when number of input variables goes high and weight vector is large in size[6]. This consumes a large memory and processing time and acts as a bottleneck in reaching towards the point of minima.

An improvement on simple Newton Method is Quasi–Newton technique in which tedious computations of hessian are replaced by an approximation for the hessian [7, 9]. The Levenberg Marquardt Algorithm (LMA) provides a trust region approach to Gauss–Newton Algorithm [12] and interpolates between Gauss–Newton Algorithm and steepest descent algorithm. However, LMA is considered more powerful than Gauss–Newton Algorithm, because of its capability to find a solution in cases, when starting point is very far off the point of minimum error.

Method of conjugate gradient (CGM) bypasses the computation of second order derivatives, and the idea is to restrict the search directions toward paths that are orthogonal to all previous searches [11]. Advantages of CGMare that they consume relatively lesser memory for large size problems and each step is quite fast [4]. It utilizes a simple line search to find the required step size in the search direction and jumps deep inside valley of error gradient. The line search avoids the tedious calculations of the hessian matrix; but needs to calculate the error gradient at a number of points in search direction.CGM saves the time for complex computation of second order derivatives but still line search along conjugate directions every time is very time consuming. The scaled conjugate gradient method (SCGM) avoids the need of time consuming line search [12, 13] and is applicable to larger networks. This method suppresses the instability in computation by combining the trust region approach of LMA with the CGM approach [12]. SCGM regulates the indefiniteness in tedious calculation of hessian with a scalar value and computes the optimal step size, without complex and costly calculations done during line search by the standard CGM.

3.0 FIRST & SECOND ORDER ALGORITHMS EMPLOYED FOR TRAINING OF NETWORK

We have applied following techniques for training the predictive neural networks.

3.1 Gradient Descent (GD) Algorithm

We move towards the point of minimum of error gradient along error surface in very small steps and descend in opposite direction of the error gradient [14]. New updated weight vector is computed as:

$$w_{ji_{next}} = w_{ji_{prev}} + \Delta w_{ji} \tag{5}$$

A control parameter η is introduced as shown in Eq. (5) to have an extra control over the speed of training which can regulate the amount of overall corrective adjustment applied to weight vector.

$$w_{ji_{next}} = w_{ji_{prev}} + \eta(T_j - O_j)I_i \tag{6}$$

But here learning rate parameter η is kept constant leading to a slow convergence towards minima. In areas where the error surface is flat and error derivative is small in magnitude and here a small value of step size is required to find a significant decrease in error. On the contrary, if the error surface is highly curved and the derivative is large in magnitude and here small values will reduce the speed of convergence. Some variations are suggested to improve in these kinds of situations [20, 23, 24, 25].

3.2 Gradient descent with adaptive learning rate (GDA)

GDA attempts to keep the learning step size large for the speedy convergence and also keeping the training stable. Learning rate is varied according to the complexity of the error surface. If the new error surpasses the old error by a predefined value, the new weight vector is rejected. Also, the rate of learning parameter is decreased by multiplying with fractional value which is less than 1. Otherwise, the new weight vector is retained. If the new error is lesser than previous error, the learning rate is further increased by multiplying with a factor slightly greater than 1 [21, 23].

Silva and Almeida implemented the same idea in a simpler way and suggested the changes to be done in constant fractional values. Rate of learning in the training algorithm is enhanced by multiplying with a fixed fractional value slightly more than 1, generally 1.20 and reduced by multiplying with the reciprocal of fixed fractional value (i.e. 1/1.20), just to increase

or decrease by a small fraction. Momentum term can also participate in the algorithm; but is kept at fixed value and is non-adaptive in this method [21]. A backtracking process is utilized to keep an eye on the continuous increase in the error values and to revert back after a number of iterations to the historical points of lesser error values as accomplished in the preceding iterations.

3.3 Gradient descent with adaptive momentum (GDM)

GDM allows the network to be sensitive not only to the error gradient, but also respond to the latest trends in the error surface. Momentum term allows the network to ignore small irregularities on the error surface. Without momentum term a network can be caught in a small narrow local minimum; and momentum helps to slide through such a minimum. The parameter momentum constant α determines the amount of influence of previous iterations on the on-going iterations and hence can be called to approximate the second order algorithms [14, 22].

Here, delta rule now becomes:

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}(n)} + \alpha \Delta w_{ji}\ (n-1)$$

The reduced learning due to momentum parameter tries to slow down the search in irregular areas on the error surface and increased learning rate enhances the search in smooth areas on the error surface [21].

In initial studies, momentum parameter was kept fixed but later studies discovered that this technique suffers from unnecessary acceleration, when the current error gradient is in opposite direction to the previous searches. This results to move in the upward direction instead of going down the slope as required. Therefore, it is necessary that the momentum to be adjusted adaptively instead of setting it to a constant value [19, 21].

3.4 Conjugate Gradient Method (CGM)

Early methods involve the calculation and storing of second order derivatives of error energy, the hessian and its inverse, which are very tedious and memory consuming, when the number of prediction variables is very large and overall weight matrix is large in size. Hence, in these situations it is better to employ algorithms like CGM; which avoids the computation of second order derivatives but still achieves quadratic termination. To avoid the need of second order derivatives, while descending across the error surface, we restrict the search direction to the paths that are orthogonal to all previous searches. An enhancement over the steepest descent method, the conjugate gradient method comprises of these computational steps:

1. Choose the initial search direction opposite to error gradient i.e. $d_0 = -g_0$.
2. Compute an optimal value for the training parameter α_k for minimum value of the gradient function with help of line search method.
$$x_{k+1} = x_k + \alpha_k d_k$$
3. Find out a new search direction orthogonal to all previous searches and compute the parameter β_k such that:
$$d_{k+1} = -g_{k+1} + \beta_k d_k$$
4. If convergence to point of minima is not reached or the stopping conditions set for training are not met, then proceed to second step of algorithm.

Researchers have developed a number of variations for calculation of parameter β_k such as Hestenes and Stiefel variation, Fletcher and Reeves variation, Polak and Ribiere variation are widely used for practical purposes. The main advantage of conjugate gradient algorithm is that it uses relatively less memory when applied to bigger networks and bypass computations for second order derivatives. But on the negative side the line search method may require more computation for an optimal value of the parameter α_k [11].

3.5 Scaled Conjugate Gradient Method (SCGM)

Similar to CGM, while computing the error gradient, SCGM bypasses the tedious and memory consuming computation of complex hessian matrix and computes an approximation which is close to second order derivatives. A new search direction d_k ; but a new step size α_k are calculated every time during k^{th}iteration in order to update the weight vector for training of network, such that:

$$E(W_k + \alpha_k d_k) < E(W_k)$$

The quadratic approximation on the error surface $E(W_k)$ in a neighboring point of the weight vector W_k is given by the Taylor's approximation as [16]:

$$E(W_k + w) \approx E(W_k) + E'(W_k)^T w + \tfrac{1}{2} w^T E''(W_k) w$$

But it is very time and memory consuming to exactly calculate the hessian $E''(W_k)$ and particularly when weight vector is large in size, therefore, second order information S_k is approximated in terms of first order derivatives as:

$$S_k = E''(W_k) d_k \approx \frac{E'(W_k + \sigma_k d_k) - E'(W_k)}{\sigma_k}$$

$$\text{for } 0 < \sigma_k \leq 1$$

Where, $E(W_k)$ denotes the multidimensional error surface and W_k is the weight matrix for k^{th}iteration, $E'(W_k)$ is the error

gradient. And parameter σ_k denotes the incremental change in the network weights for this second order method. In this algorithm, the trust region approach of Levenberg Marquardt algorithm is used with the conjugate gradient approach to calculate the next step size [12]. Variable c_k is used to control the indefiniteness of $E''(W_k)$. This is done by computing the second order information as:

$$S_k = \frac{E'(W_k + \sigma_k d_k) - E'(W_k)}{\sigma_k} + c_k d_k$$

And at every step, we keep on computing value $\delta_k = {d_k}^T S_k$ for checking the indefiniteness of $E''(W_k)$. We keep on adjusting c_k and keep looking at the sign of δ_k in each iteration, which checks that hessian $E''(W_k)$ is not positive definite. When $\delta_k \leq 0$; c_k is increased and S_k is calculated again. The new weight vector is now calculated as:

$$W_{k+1} = W_k + \alpha_k d_k$$

, till we reach the point of minima or stopping conditions for the training are met.

4.0 EXPERIMENTAL OBSERVATIONS AND RESULTS

Training functions of first and second order (traingd, traingda, traingdm, traincgp, trainscg) available in MATLAB Neural toolbox package were employed to train the neural networks. Large data sets taken from live data warehouse are employed for experimentation and model development. Training performance based on Mean Squared Error (MSE), gradient plot for algorithm convergence are observed to check for the behavior and efficiency of gradient algorithms under consideration. A variety of neural networks architectures are tested for development of desired prediction models. But two layered architecture with 15 neurons is observed as the optimal configuration with data sets employed under similar hardware/software configurations.

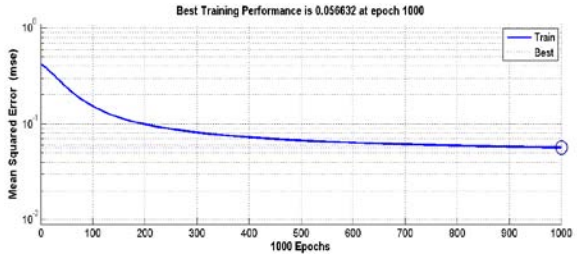


Figure 1: Performance plot when employing simple gradient descent learning

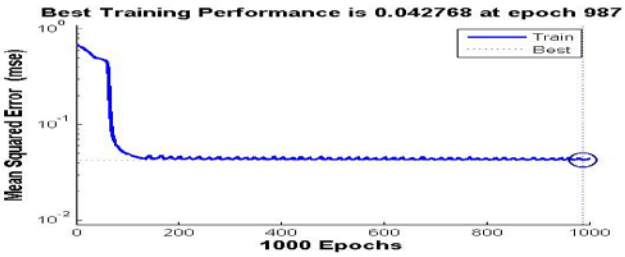


Figure 2: Performance plot when employing gradient descent with adaptive learning

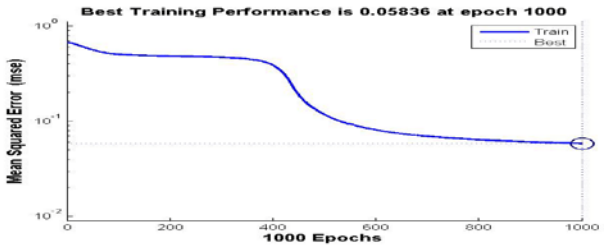


Figure 3: Performance plot when employing gradient descent with adaptive momentum

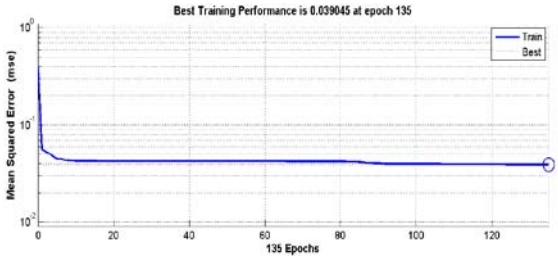


Figure 4: Performance plot when employing conjugate gradient learning

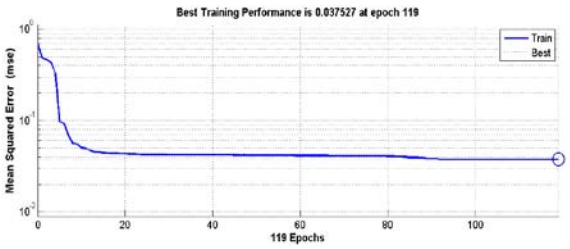


Figure 5: Performance plot when employing scaled conjugate gradient learning

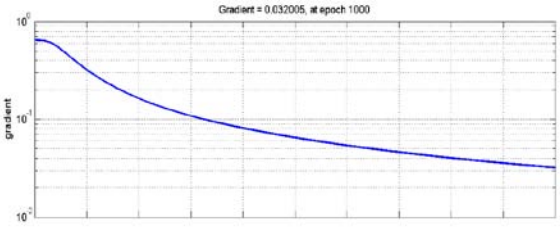


Figure 6: Error gradient plot when employing simple gradient descent learning

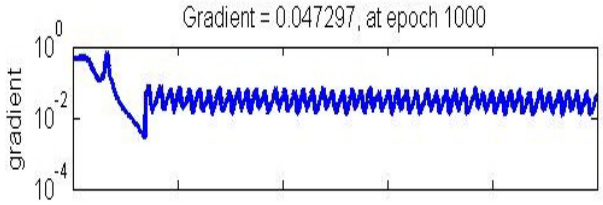


Figure 7: Error gradient plot when employing gradient descent with adaptive learning

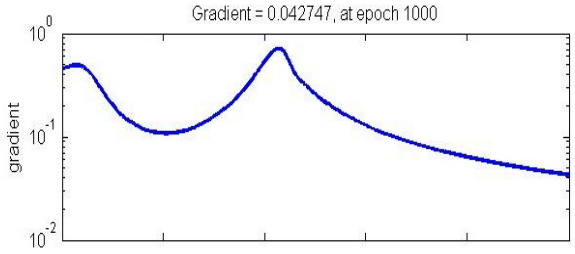


Figure 8: Error gradient plot when employing gradient descent with adaptive momentum

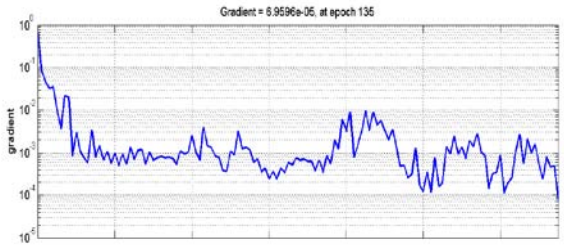


Figure 9: Error gradient plot when employing conjugate gradient learning

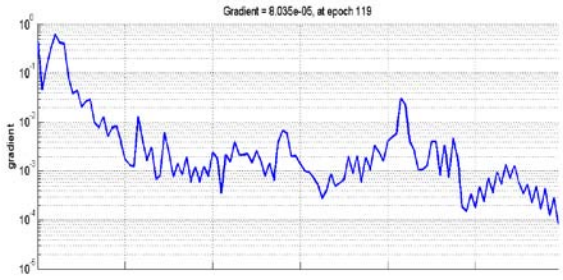


Figure 10: Error gradient plot when employing scaled conjugate gradient learning

Figures 1 to 5 show changes in mean square error (MSE) verses numbers of epochs achieved in the best cases while applying the simple, adaptive(learning, momentum) and second order gradient methods. Error gradient plots, while employing different training methods are shown in Figures 6 to 10. It has been observed that gradient decent, adaptive learning and adaptive momentum methods were unable to accomplish the error gradient of 10^{-4} even in the 1000^{th} epoch but second order methods like CGM converged in 135 epochs and SCGM in 119 epochs.

Table 1: Experimental results of employing different gradient based learning algorithms of first and second order

Training Method	Steepest (Gradient) Descent	Gradient descent with adaptive learning rate	Gradient descent with momentum	Conjugate Gradient (Pollok Variation)	Scaled Conjugate Gradient
Training Function	traingd	traingda	traingdm	traincgp	trainscg
Hidden Layer Neurons	15	15	15	15	15
MinimumGradient	0.0001	0.0001	0.0001	0.0001	0.0001
TransferFunction	TANSIG	TANSIG	TANSIG	TANSIG	TANSIG
Final No. of Epochs	10^3	10^3	10^3	135	119
TrainingTime	0:16:14	0:14:39	0:15:27	0:07:24	0:04:41
TrainingPerformance	0.0566	0.0427	0.05836	0.0390	0.0375
InitialGradient Value	0.6460	0.4470	0.4470	0.6760	0.4470
FinalGradient Value	3.20E-02	0.0473	0.0427	6.96E-05	8.04E-05

5.0 CONCLUSION

The main focus of the research is to improve the speed and accuracy of convergence in network training. Convergence of adaptive methods is compared with simple gradient descend and second order methods. It is concluded that adaptive gradient based method are slightly better than simple gradient but still there performance is not comparable to second order techniques. Adaptive methods took lesser training time than simple gradient but they were not able to converge even in 1000 epochs toward error gradient of the order of 10^{-4}. While on the other hand, second order techniques were able to reach an accuracy level of 10^{-4} and 10^{-5} and prove far better in terms of training time and accuracy. Simple gradient (GD) with constant learning rate has shown the poor convergence, conjugate and scaled conjugate gradient methods (CGM, SCGM) show the fastest convergence and adaptive methods (GDA, GDM) fall in between, in terms of convergence towards the minimum error gradient.

REFERENCES

[1]. S. V. Chande and M. Sinha, "Genetic Algorithm: A Versatile Optimization Tool", *BIJIT - BVICAM's International Journal of Information Technology*, Vol. 1, No. 1, pp. 7-12, Jan.-Jun. 2009.

[2]. J. C. Meza, "Steepest descent", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2, No. 6, pp. 719-722, Dec. 2010.

[3]. S. Osowski, P. Bojarczak, and M. Stodolski, "Fast second order learning algorithms for feed forward multilayer neural networks and its applications", *Neural Networks, Elsevier*, Vol. 9, No. 9, pp. 1583–1596, Dec. 1996.

[4]. R. Battiti, "First & second order methods for learning between steepest-descent & Newton's method", MIT Press, *Neural Computation*,vol. 4(2), pp. 141-166, Mar. 1992.

[5]. A. K. Verma, R. Anil and Dr. O. P. Jain, "Fuzzy Logic Based Revised Defect Rating for Software Lifecycle Performance Prediction Using GMR", *BIJIT - BVICAM's International Journal of Information Technology*, Vol. 1, No. 1, pp. 1-6, Jan.-Jun. 2009.

[6]. R. Rojas, *Neural Networks - A Systematic Introduction*, Berlin, Springer, 1996.

[7]. A. Likas, and A. Stafylopatis, "Training the random neural network using Quasi–Newton methods", *Euro. Jour. of Operational Research, Elsevier*, Vol. 126, Issue 2, pp. 331-339, Oct. 2000.

[8]. R. Rastogi, S. Agarwal, P. Sharma, U. Kaul and S. Jain, "Business Analysis and Decision Making Through Unsupervised Classification of Mixed Data Type of Attributes Through Genetic Algorithm", *BIJIT - BVICAM's International Journal of Information Technology*, Vol. 6, No. 1, pp. 683-689, Jan.-Jun. 2014.

[9]. S. M. A. Burney, T. A. Jilani, and C. Ardil, "A Comparison of first and second order training algorithms for artificial neural networks", *International Journal of Computational Intelligence*, Vol. 1, No. 3, pp. 176-182, 2005.

[10]. M. T. Hagan, and M. B. Menhaj, "Training feed forward networks with Marquardt algorithm", *IEEE Tran. on Neural Networks*, Vol. 5, No. 6, pp. 989-993, Nov. 1994.

[11]. E. K. P. Chong, and S. H. Zak, *An Introduction to Optimization*, 2nd Edition, John Wiley and Sons, 2001.

[12]. M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning", *Neural Networks, Elsevier*, Vol. 6, Issue 4, pp. 525–533, 1993.

[13]. J. Lunden, and V. Koivunen, "Scaled conjugate gradient method for radar pulse modulation estimation", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Proceedings ICASSP'07, Vol. 2, pp. 297–300, Apr.2007.

[14]. M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*, Chapter 12, PWS Publishing, Boston, 1996.

[15]. S. Goel, J. B. Singh and A. K. Sinha, "Traffic Generation Model For Delhi Urban Area Using Artificial Neural Network", *BIJIT - BVICAM's International Journal of Information Technology*, Vol. 2, No. 2, pp. 239-244, Jul.-Dec. 2010.

[16]. Swanston D. J., Bishop J. M., and Mitchell R. J., "Simple Adaptive Momentum New Algorithm for training Multilayer Perceptron, *J. Engineering Letters*, Vol. 30 (18), pp. 1498-1500, 1994.

[17]. G. Kumar and P. K. Bhatia, "Optimization of Component Based Software Engineering Model Using Neural Network", *BIJIT - BVICAM's International Journal of Information Technology*, Vol. 6, No. 2, pp. 732-742, Jul.-Dec. 2014.

[18]. Y. H. Zweiri, L. D. Seneviratne, and K. Althoefer, "Stability Analysis of a Threeterm Back Propagation Algorithm", *Neural Networks,Elsevier,*vol. 18, no. 10, 2005.

[19]. M. Z. Rehman, and N. M. Nawi, "Studying the Effect of Adaptive Momentum in Improving the Accuracy of Gradient Descent Back Propagation Algorithm on Classification Problems", *International Journal of Modern Physics, World Scientific,*Vol. 1(1),pp. 1–5, 2010.

[20]. R. A. Jacobs, "Increased Rate of Convergence through Learning-Rate Adaptation", *Elsevier, Neural Networks*; 1:295–307, 1988.

[21]. M. Moreira, and E. Fiesler, *Technical-Report IDIAP, Neural Networks with Adaptive Learning Rate & Momentum Terms*, No.95-04, Oct.1995.

[22]. Y. Bai, H. Zhang , and Y. Hao, "The performance of the back propagation algorithm with varying slope of the activation function", *Chaos, Solitons and Fractals, Elsevier*, 40, pp.69–77, 2009.

[23]. Saduf, M. A. Wani, "Comparative Study of Back Propagation Learning Algorithms for Neural Networks", International *Journal of Advanced Research in*

Computer Science and Software Engineering, Volume 3(12), pp. 1151-1156, December 2013.

[24]. M. Z. Rehman, and N. M. Nawi, "Improving the Accuracy of Gradient Descent Back Propagation Algorithm on Classification Problems", *Int. Journal on New Computer Architectures and Their Applications*,1(4): pp. 838-847, 2011.

[25]. G. D. Magoulas, M. N. Vrahatis, and G. S. Androulakis, "Improving Convergence of the Back-Propagation Algorithm Using Learning-Rate Adaptation Methods", MIT Press, *Neural Computation*,vol. 11, No. 7,pp. 1769–1796 Oct. 1999.

Hindrances in Providing e-Commerce Services in Saudi Retailing Organizations: Some Preliminary Findings

Abdullah Basahel[1] and **Kamel Khoualdi**[2]

Submitted in February 2015; Accepted in April, 2015

Abstract - e-commerce, especially e-retail, has changed the concept of marketing and business in developed countries, where the customer does not need to go to the market to get the goods. This process is not yet successful in Arabian countries, particularly Saudi Arabia. This study aims to identify the most important obstacles that prevent enterprises from providing electronic sales to their customers. I found that 86% of facilities do not provide these services. The most important obstacles that prevented them doing so are the inefficient laws and electronic payment systems, in addition to the lack of awareness and confidence of consumers in these services.Upon conducting a survey, we estimate that 86% of businesses do not provide these services. The biggest obstacles in the provision of these services are the inefficient and deficient laws and inapt electronic payment systems. In addition there is a lack of awareness and confidence amongst consumers in relation to these services

Index Terms – E-Commerce, Retailers, e-Tailers, Saudi Arabia.

1.0 INTRODUCTION

Information technology is no longer just a tool to use but has become a work environment characterized by rapid and successive developments and competitive advantages. The start of the revolution of information technology is electronic data exchange (EDI) within organizations and between them as an initial step to entering the field of e-commerce. The significant growth in the volume of international business with globalization encourages organizations to use e-commerce to sell their products and makes countries keen to develop laws for these new styles of trade patterns. The term "e-commerce" is related to terms such as "e-business," "e-management," and "e-market," and with other concepts produced by the digital economy of the Internet and the information base and networks.

E-commerce, and particularly electronic sales and purchase, is no longer a luxury as far as it becomes a necessity for the society as a result of the openness to trade of developed countries trade which offer these kind of services. In any commercial transaction, there are two main parties: business organizations, in particular retailers, and consumers. Undoubtedly, each of them desires to excel in the provision of services but both of them encounter considerable obstacles in the way of adoption and use of e-commerce. In the research leading to this article, we were mainly concerned with finding revolution, face the fierce competition and meet the needs of the community. E-commerce, which means business trading.

the kind of constraints that companies are facing in the way of adoption and usage of e-commerce in Saudi Arabia.

The twenty first century is the era of the Internet economy, and to be a part of this era, it becomes imperative for organizations to adopt e-commerce in order to cope with the new electronic through Internet, can have benefits for organizations as well as consumers. From the business perspectives, more transactions can be accomplished resulting in an increase in sale and cost reduction. As for the consumer, e-commerce will provide him with a better service, speed in shopping, ability to compare between products and ease in taking a purchase decision. The importance of research is an attempt to review some of the available solutions to avoid obstacles of adopting e-commerce by organizations and consumers in order to increase the benefits for both parties.

Businesses strive for increased revenue. For achieving this crucial objective, they expand and create innovations in their businesses. Obviously e-Commerce is a new paradigm which businesses find very cost effective to provide. So, the business can raise much needed revenue by spending relatively small amount of money as e-Coomerce economizes on resources such as infrastructure and workforce. So, more and more revenue is generated as a result of customer transactions of e-Commerce. On the other hand, impressed transactions result in reductions of costs by way of the economy of scale. As for the consumer, e-commerce will provide them with a hassle free and better service, reduction if time spent in e-shopping, increased ability to compare between products for making an informed decision. As an important step in our research is an attempt to review some of the available solutions to avoid prevailing obstacles of adopting e-commerce by organizations and consumers in order to increase the benefits for both parties. In our research, we have chosen the case of a well-known Middle Eastern and Arabian country namely, Saudi Arabia

2.0 BACKGROUND

The rapid development of technology, and particularly the unveil of Internet in the beginning of the nineties, has contributed to the emersion of e-commerce, which offers many business opportunities and hence having a great impact on countries' economic future development and international competitiveness. The amount of goods and services purchased online has increased considerably. The last ten years has seen a

[1,2]*Department of management Information Systems, King Abdulaziz University, Jeddah 21589, Saudi arabia*
E-mail: [1]abasahl@kau.edu.sa and [2]kkhoualdi@kau.edu.sa

proliferation in B2C e-commerce, which is rapidly changing. For this reason, there are opportunities for, retailers and researchers to better understand how and why customers are involved in e-commerce and continue to mutually interact. Online shopping is a growing trend all over the world, as is the same in the Kingdom of Saudi Arabia, since online shopping is a very easy and convenient way to buy.

Surprisingly, there is a lack of Arabian studies, especially Saudi studies, compared to research and studies in developed countries. Of the Arab research studies on the e-commerce, most were concerned with the electronic trading between businesses (B2B) instead of between businesses and customers (B2C). By contrast, the electronic libraries of developed countries are filled with studies, articles, and reports related to all aspects of e-commerce, from obstacles and solutions, customer satisfaction and the extent of their loyalty, to e-shops and consumer behavior on websites. The review of the literature allowed us to formulate questions and identify the most important obstacles behind the lack of e-commerce in KSA.

This is the age of communication. The numbers of Internet users around the world grew to 1,966,514,816 in 2010. The concept of electronic commerce (e-commerce) has grown accordingly.

There are multiple definitions of e-commerce due to the multifaceted concept. E-commerce can be defined as the purchase or sale of goods and services over computer networks[1] and [2].

Increase in the supply of goods and services worldwide have pushed organizations for adapting e-commerce business models, consequently removing the barriers between companies, supplier, and customers. There are several benefits of e-commerce, including reduced cost and time, and achieving higher returns. E-commerce also provides the opportunity for small and medium-sized organizations to enter new markets and it thus increasing competitiveness. E–commerce increase selling and buying process, decrease cost of operations and facilitate communication between buyers and sellers. There are also disadvantages and constraints, like the absence of legal frameworks governing e-transactions and the difficulty of providing worker safety, confidentiality, and privacy.

The most important step in the stages of e-commerce is electronic payment, often involving e-payment methods such as credit card, smart cards, and e-cash. This is the most dangerous stage, and it is very important to consider safety in the operations of e-payment and compliance with all laws designed for e-commerce. These laws can be difficult to apply because of the lack of geographical boundaries on the Internet.

2.1 E-retail

The term "e-retail"[3]is defined as doing business directly with the individuals buying and selling through the World Wide Web (Internet), taking into account the fact that buying and selling here includes goods, services and intellectual property rights. There has been large recent interest in electronic retailing because the commercial use of the Internet by e-retailers has witnessed steady growth: e-retailing in the USA in 2008 was $142 billion, approximately 4% of the country's total retail sales.

E-retail constitutes a small, but growing portion of retail activities [4]. Competing in this dynamic and technologically complex retail environment is a challenge for retail companies which may have broad impact on their organizational and spatial structure, as well as shopping models. Diversifying their activities by adding an online model to their existing business models consisting of stores, catalogs, and direct marketing seems to be indispensable to keep up with the competition.

E-retailing gives customers options that are not available in traditional trade, such as rapidly moving from one location to another, reviewing commodity profiles, and comparing prices and quality between goods all at once. Even consumer behavior is different in the procurement process via the Internet. Some of the elements that affect the decisions in the e-shopping include the buyer's personality, age, gender, and culture.

There are four types of traders who predominate in the e-retail market: traders versed in the real world, visual, brokers, and manufacturers. Most successful traders in this area have places or markets in the real world. The reason for their success is because they have experience in the field of trade and trading in general.

2.2 E-Commerce in Saudi Arabia

In 1999 a royal order established the Permanent Technical Committee of E-commerce to monitor developments and identify the needs of the communication infrastructure and technical requirements, security, legal requirements, and the Commission is composed of a group of government agencies. The use of information technology in the Kingdom of Saudi Arabia had a major shift from 2001 to 2009, with a four-fold increase in spending on information technology products and services by the government, corporations, and individuals; the annual growth rate was 18.5%. This made the KSA the largest spender on communications and information technology in the Middle East.

Because of this interest in IT, the number of Internet users in the kingdom is increasing at the individual and commercial levels. The number of Internet users in the kingdom was 5% in 2001, and then reached 41% in 2010 . The kingdom ranked as 38th in 2009-2010 in terms of e-readiness, while it was ranked 45 in 2003. The Saudi Post has recently launched E-Mall site that specializes in selling online. Also, the banking sector in KSA is a pioneer in e-commerce and has established all e-banking services for their customers. Since e-Commerce is a byproduct of internet, therefore, the health of e-commerce can be determined to an extent by looking at the usage of the internet. Figure 1 (http://www.tradingeconomics.com/saudi-arabia/internet-users-wb-data.html)shows the number of internet users in Saudi Arabia which has been growing steadily. According to a daily newspaper of Saudi Arabia, Arab News (http://www.arabnews.com/news/463005), the value of the Saudi Arabian e-Commerce is about fifteen billion dollars annually.

The Kingdom of Saudi Arabia (KSA) has registered a record overall growth of 43% in e-Commerce in Q1 of 2014 compared to the same period last year, making it the highest growth rate in the MENA region, according to estimates by Visa, one of the world's largest retail electronic payments networks. For details, see [5]. However, despite all the development and growth in IT in the kingdom, there are still challenges and obstacles [6], such as a lack of human skills. E-commerce still has a low level of implementation, especially in retail sectors. The use of e-commerce is primarily confined to large enterprises such as SABIC and Saudi Aramco in their e-transactions (B2B). But there is a dearth of sites that offer goods for consumers, despite the growth of the traditional retail sector in the kingdom, where reports predict the annual growth of major stores and shops, by 4.4% and 7.2%, and up to 24.2 billion RS and 18.7 billion, respectively.

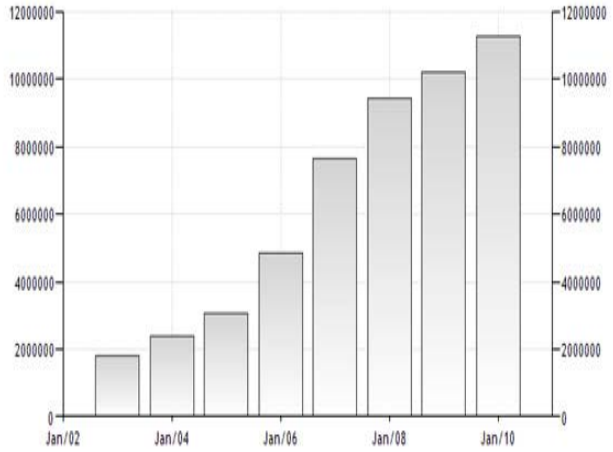


Figure 1: Internet users in Saudi Arabia

Technological, social and economic activates have a natural link with E-commerce. Therefore, it may be useful to mention some details of studies already conducted in these areas. Indeed there are a number of studies available on various aspects of Saudi Arabia. For example, Mohammad Yamin [7] and [8] has provided an insight into the annual pilgrimage known as Hajj and associated issues likehealth, safetyand economic activities of Saudi Arabia. Basahel [9] has provided details of some educational aspects of the women in Saudi Arabia. Yamin and Huddoff [10], and Yamin and Makrami[11] have researched various technological aspects and their impacts into Saudi Arabian society and economy.

3.0 RESEARCH METHOD

The study is based on the descriptive analytical approach in a manner consistent with the nature and goal of the problem. Data was collected using a questionnaire containing questions related to the problem and then distributed to the retailers in Saudi Arabia. The study sample is composed of 50 retailers in different activity domains in different regions in Saudi Arabia.

The purpose of the present study was to assess the difficulties and obstacles retailers are facing to adopt e-commerce. In the present study, we addressed four research questions:

1) Does the lack of supporting technologies inhibit providing e-commerce service in KSA?
2) Does the weakness of laws and regulations impede the provision of e-commerce service?

3) Does the nature of the Saudi consumer hinder the provision of e-commerce service?
4) Does the capacity of the facilities hamper e-commerce services?

Each of the previous questions had been provided with three options in the survey. Table 1 shows the responses of the sample.

Table 1

Obstacles	Average	Standard Deviation	Degree ofapproval
Limitation of electronic payment systems in the Kingdom of Saudi Arabia	3.63	1.474	**Agree**
The inefficiency of the current laws of electronic selling	3.63	1.159	**Agree**
Lack of support from government agencies related to issues of electronic sale	3.60	1.329	**Agree**
Lack of trust of consumers to purchase from the Internet	3.57	1.251	**Agree**
Electronic signatures/seals are not widespread in the region.	3.53	1.332	**Agree**
There are no laws and rules to organize the electronic trade.	3.47	1.167	**Agree**
A lack of awareness among Saudi Arabia's consumers about purchases via the Internet	3.47	1.167	**Agree**
Difficult to acquire credit cards for all consumers to make electronic payments	3.33	1.422	**Neutral**
Most consumers do not have not a mailing address, which will hinder the delivery process.	3.33	1.493	**Neutral**
Poor telecommunications infrastructure in the kingdom prevents the company from selling online.	2.87	1.383	**Neutral**

The difficulty of providing confidentiality and safety factors in electronic payment	2.87	1.502	**Neutral**
The lack of specialized staff and expert employees in the IT field in the company	2.80	1.270	**Disagree**
Refusal from senior management to activate this function for consumers	2.70	1.149	**Disagree**
The number of Internet users in the kingdom is too small for the company to provide this service.	2.60	1.221	**Disagree**
The nature of the product prohibits its sale via the Internet.	2.47	0.999	**Disagree**
The size of the company does not allow it to offer electronic sales.	2.43	1.145	**Disagree**

As shown in the table, the limitation of electronic payment systems and the inefficiency of the current laws were most cited as the obstacles and challenges that hinder the sale of products electronically through the Internet (M = 3.63 for each).

When I asked enterprises to mention some of the methods to develop e-commerce, their answers revolved around the following:

- Activating the laws relating to e-commerce, to protect all parties of the process.
- Improving the awareness of the consumer and businesses about the importance of electronic commerce.
- Facilitating the procurement methods through diversification to make electronic payments to be available to all.
- Activating the concept of an electronic signature or seal for orders.
- Numbering buildings, blocks, and streets to ease the delivery process, perhaps introducing a ZIP code system.

4.0 CONCLUSIONS

The literature review and the results of this study revealed several ways to reduce the impact of current constraints on e-commerce in the Kingdom of Saudi Arabia, as follows:

- Those responsible for e-commerce in the Kingdom should enact clear and explicit laws regarding e-transactions to ensure the rights of all parties; these laws should benefit from the experiences of other countries that have preceded the KSA in this area.
- The Chambers of Commerce administration should raise the awareness of the importance, advantages, and benefits of electronic retail services to encourage facilities to take this step.
- The Ministry of Trade and Chambers of Commerce should support the businesses to activate these kinds of services, and encourage and create collaborations and alliances between the parties that could contribute to the development of this aspect.
- Saudi Arabian Monetary agency (SAMA) should cooperate with the banks to introduce electronic payment systems accessible to all segments of society, so as to ensure the ability of customers and businesses to complete this type of purchase.
- Other agencies should emulate Saudi Postal and develop e-malls and similar projects or portals to increase and spread e-services in Saudi society.

REFERENCES

[1]. M. Bui, and E. Kemp. 2013. "E-Tail Emotion Regulation: Examining Online Hedonic Product Purchases". International Journal of Retail & Distribution Management, Vol. 4, No. 2, pp. 155 – 170, 2013

[2]. Y. Kim, and G. Li, "Customer Satisfaction with and Loyalty Towards Online Travel Products". Tourism Economics.Vol. 15, pp. 825 – 846, 2009.

[3]. O. M. B. Kolesar, and R.W. Galbraith, "A services-marketing perspective on E-retailing: implications for E-retailers and directions for further research", Internet Research-Electronic Networking, Vol. 10, No. 5, pp. 424–438, 2000.

[4]. O. Rotem-Mindali, and I. Salomon, "The impacts of E-retail on the choice of shopping trips and delivery: Some preliminary findings",Transportation Research Part A: Policy and Practice, Vol. 41, No. 2, pp. 176–189, February2007.

[5]. ZAWYA, eCommerce sees record growth of 43% in Saudi Arabia, Online (2014), Available from http://www.zawya.com/story/eCommerce_sees_record_growth_of_43_in_Saudi_Arabia-ZAWYA20140612055548/

[6]. M. R. Noruzi. "E-Commerce in Iran: Barriers and Suggestions" Interdisciplinary Journal of Contemporary Research In Business. Vol. 2, No. 7, November 2010.

[7]. Mohammad Yamin, Health Management in Crowded Events: Hajj and Kumbh, BIJIT - BVICAM's International Journal of Information Technology, January - June, 2015; Vol. 7 No. 1; ISSN 0973 – 5658

[8]. Mohammad Yamin, Cloud Economy of Developing Countries, World Journal of Social Sciences, Vol. 3. No. 3. May 2013, Pp. 132 – 142

[9]. Abdullah Basahel, An Empirical Study of Impacts of ELearning in Female Higher Education in Saudi Arabia, BIJIT - BVICAM's International Journal of Information Technology, January - June, 2015; Vol. 7 No. 1; ISSN 0973 – 5658

[10]. Mohammad Yamin and Abdullah A. Al Hudhaif, MIS in Small Industry: Sanitary Ware in Saudi Arabia, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN

2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 11, November 2014)

[11]. Mohammad Yamin and Ammar A. Al Makrami, Cloud Computing in SMEs: Case of Saudi Arabia, BIJIT - BVICAM's International Journal of Information Technology, January - June, 2015; Vol. 7 No. 1; ISSN 0973 – 5658, 853-860.

Feature Extraction of Voice Segments Using Cepstral Analysis for Voice Regeneration

P. S. Banerjee[1], Baisakhi Chakraborty[2] and **Jaya Banerjee[3]**

Submitted in February, 2015; Accepted in April, 2015

Abstract–Even though a lot of work has been done on areas of speech to text and vice versa or voice detection or similarity analysis of two voice samples but very less emphasis has be given to voice regeneration. General algorithms for distinct voice checking for two voice sources paved way for our endeavor in reconstructing the voice from the source voice samples provided. By utilizing these algorithms and putting further stress on the feature extraction part we tried to fabricate the source voice with different pitch and intonation patterns. This process of uniquely tracing the features and re assembling them to reproduce the target voice is what we have concentrated on in this work. While doing so the aspect of liftering and cepstrum analysis has been utilized to the fullest.

Index Terms – Cepstral Analysi, liftering, cepstrum

1.0 INTRODUCTION

Voice Processing or Speech Processing may be regarded as one of the most integral part of any module dedicated for either speech recognition, as an authentication for granting unique entity access or for speech to text or vice versa. Linguistically and phonologically we may not be able to decipher much from the available waveform due to enormous amount of data for each audio signal that is processed, where as if interpreted in a more sophisticated manner then we may be able to extract relevant unique information for every new incoming data. [1] While defining a cepstrum we can say that it is the output of the estimated spectrum's logarithm is computed and its Inverse Fourier transform is also calculated. The different categories under which the cepstrum may be divided is a complex cepstrum, a real cepstrum, a power cepstrum, and phase cepstrum. Power cepstrum finds its application in the analysis of human voice. As a baseline fact the term "cepstrum" basically originates from the word "spectrum" and is achieved by just reversing the first four letters. The fundamental procedures of cepstra may be listed as quefrency analysis, liftering, or cepstral analysis. The rate at which the various spectrum bands change may be called as cepstrum. Out of the many utilities of it the most primitive use of it was for characterizing the seismic echoes resulting from earthquakes and bomb explosions. But to the best of our concern it is also used to determine the fundamental frequency of human speech. Cepstrum pitch determination is particularly effective because the effects of the vocal excitation (pitch) and vocal tract (formants) are additive in the logarithm of the power spectrum and thus clearly separate. The best possible step will be to have smaller amount or relevant data which may act as a building block for the voice to be recreated. At the outset we can imagine that any signal emanating from a source is a cumulative result of the provided file and the basic response to the file in coordination to it. Both the input as well as the output component are to be treated separately for further processing and this process may be mathematically defined as the deconvolution process.[18]

There are algorithms like Linear predictive coding (LPC), Hybrid Harmonic/Stochastic (HYBRID H/S) and last but not the least the TD PSOLA which we can use to calculate the perturbation or the excitation.[2]

1.1 Organization of the paper

This whole work comprises of the the basic problems associated with the feature extraction of the voice samples. For feature extraction the various techniques available are discussed with their relative advantages and dis advantages. The problem definition has been clearly emphasized and the cepstral analysis part has been forwarded as the possible alternative.

2.0 PROBLEM DEFINITION

At the outset we can imagine that any signal emanating from a source is a cumulative result of the input excitation and the basic system response in coordination to it. Talking in the context of mathematics and digital signal processing the convolution of the input signal with the response of the system may be regarded as the actual output. Bothe the input as well as the output component are to be treated separately for further processing and this process may be mathematically defined as the deconvolution process.[38] Attempts have been made to extract the features of the audio signals more efficiently. If we have the extracted features then mathematically it should be possible to to uniquely identify a speech signal in its digitized form. But due to the massive diversity of the speech waveform and vast amount of data relative to a particular speech waveform we will always require a huge amount of information to be stored for a particular uttered phonetics. Due to this very hindrance it becomes difficult to recreate back the same

[1]*Department of Computer Science & Engineering Jaypee University of Engineering & Technology, Guna (MP) 473 226, India*
[2]*Department of Information Technology, NIT Durgapur West Bengal, India*
[3]*Aryabhatta Institute of Engineering and Management, Durgapur, India*
E-mail:[1]partha1010@gmail.com,
[2]baisakhi.chakraborty@it.nitdgp.ac.in and
[3]jaya2008.banerjee@gmail.com

waveform from the extracted features. [19] Our main aim would be concentrated on grouping the information available from each sample of audio signal into smaller parameters or features. There are many feature extraction techniques that are available, some of the important one are as listed as: FFT, LPC, PLP etc. [35] After the feature extraction has been done the reconstruction of the voice can be achieved by the use of the unique feature vectors.

3.0 REVIEW OF LITERATURE AND RELATED WORK DONE SO FAR

3.1 The voice reconstruction process may be divided into a sequence of steps:

3.1.1. The initial step may be regarded as the pre requisite step where the most unique features of both the source as well as the target dataset is traced and this phase is called the Analysis step.

3.1.2. In the second step we try to map the features computed in the previous step of that of the 3source voice to that of the "to be achieved voice" to as close a proximity as possible.

3.1.3. Synthesis: Last but not the least is the final phase where the modified parameters are used to synthesis or reconstruct the new speech which generally does have the target voice as well as the required prosody too if the module assists.

3.2 Voice processing

The researches on the crux topics of voice and speech is basically revolving around the paradoxical axis of speech or voice synthesis with application areas in the form of text to voice and vice versa with stress on characterization and bifurcation of two or more voice samples. Some of the broad application areas are as enumerated. [3-4]

Generally the combination issues of these are esoteric hence [5] tries to separate the basic glottal source spectra and vocal tract using glottal inverse filtering.

3.3. Review of Application Speech Synthesis Technology

The process of synthesis of speech may be classified as restricted and unrestricted when it is messaging and text-to-speech conversion respectively. These are useful for announcements and also for the people who are impaired visually.[2] In the sections to follow we are going to exemplify some of the topics.

3.3.1. Text-to-Phonetic Conversion

The basic problematic area encountered by any TTS system is linguistic to text and vice versa conversion. The process basically begins with the preprocessing of the text to be converted and then an in-depth analysis of the data for a unique and correct pronunciation. The last step involves the proper computation of the prosodic features. A few of the important steps have been discussed here.

3.3.2. Text preprocessing

The first step i.e. text preprocessing can be understood about its difficulty level from the following example where let's say any English numeral 121 may be at first read as one hundred and twenty one and 2014 as twenty fourteen if inferred as year or two thousand and fourteen for quantizing something for measurement. Some of the similar cases are the distinction between the any numeral and then stating pilot or people. The final area is the fractions and dates which are equally troublesome. 4/14 is four-fourteenths or April Fourteenth.

3.3.3. Pronunciation

The next important task is of correct pronunciation different areas in the text. There are certain types of words that bear a same spelling but have different meaning and sometimes different pronunciations also which are called homographs. Now these type of words are a big obstacle for the overall module.

3.3.4. Prosody

The rhetorical flow of any word or its segment will definitely consider the correct intonation, proper stress at the punctuations. The basic meaning of the uttered phrase and the emotional state of the speaker are some of the deciding factors for the prosodic characteristics. The dependencies of prosodic variations are shown in Figure 1. The ironic situation is that any information in written or textual format doesn't carry the traits of these qualities.[2]

A speech or voice processing system performs two types of functions which is modeling of the signal and the second one is the comparison of the features. [19]. The first stage corresponds to transformation of the signal to parameters and the second stage is the comparison of the similar features from the memory. In this regard hence forth we will be enlisting an analysis part of the related work study and then we will propose our work. Hence this part will be basically divided into

3.4. Feature Extraction:

Feature Extraction and Voice construction based on Pitch Synchronization Over-Lap Add (PSOLA) algorithm

As our study says that in this approach static voice conversion is basically taken care of. The parameters that cannot be changed by the speaker even if the person wants are called the Static parameters. They are vocal tract structure modification and these are basically inherent and natural.[8] The quantitative analysis of the algorithm is dependent on the Quality factor (Q) and The Resemblance factor (R). The parameters are applied for diversified sample of voice.

Flow of Implementation of Psola [9]

a) Application of silence removing algorithm for a silence free given input is the first step.

b) In the next step the voiced and unvoiced decision making algorithm takes care of the output of the previous step.

c) The voiced and unvoiced decision making algorithm is reutilized retrieving the qualitative data about the pitch.

d) The algorithm thus ends up right selection about the voiced and unvoiced aspect.

e) The qualitative data about the pitch is vital as markers for the pitch in the signal.

The PSOLA algorithm assesses these indexes on the pitch for scaling down of each unique signal.

f) The flow process at first checks the scale of the pitch and also converts the speech too. When we do have the data about the pitch indexes of both target as well as the source segments

then the mapping is performed for both and hence the conversion process continues. The conversion of the source to the target is achieved when the converted pitch indexes are processed through the PSOLA.
There are some other algorithms too like Linear predictive coding (LPC), Hybrid Harmonic/Stochastic (HYBRID H/S) and TD-PSOLA which we are not discussing since they are not that beneficial in the present norms.[10]

4.0 PLAN OF WORK, METHODOLOGY AND PROGRESS OF OUR OWN WORK

4.1 Our basic strategy for the module will follow the following steps: Generation of Speech Corpus or Sample (By recording the speech signal using microphone), Classification of Speech, Feature Generation, Feature Selection or Extraction, Recognition of a particular speech, Regeneration of Speech for various Prosodies.

4.2 The generation of the speech may be taken care by using condensational techniques.

4.3 Characterization or classification of speech as voiced and unvoiced and also as isolated words and connected words and sometimes as continuous speech and spontaneous speech.

4.3.1. Voiced and Unvoiced: Voiced or Unvoiced is a term used in phonetics segregate speech sounds, as either voiceless (unvoiced) or voiced. Voicing is defined as a voice when the vocal cords vibrate and is used to define phones [20]. In articulatory level, when the vocal cords vibrate a voiced sound is the output, where as a voiceless sound is one when the vocal cords do not vibrate. Its example is in the voicing of “s” and “z”.
4.3.2. Isolated word are the words which to be recognized requires each utterance to have noiseless which is disturbance less signal of the sample. Another name that might be applicable for this class is Isolated Utterance.[21]
4.3.3. The concept of Connected words can be understood as the utterances that are joined and may be regarded as separated words, but with separate utterances to be allowed with small amount of time difference between them.
4.3.4. In Continuous speech the speaker speaks in his or her natural speed and tone as a continuous voice sample. The continuous speech recognition is the most complex one because it is highly difficult to figure out the boundary of the speech.
4.3.5 The Spontaneous speech is the most natural one and is almost extempore. [22-23]

4.4 Feature Generation: For feature extraction of the speech we have employed OpenSMILE. The acoustic features in the form of low level descriptors are employed to have the values of intensity loudness and speech. [24-25].

4.5 Feature extraction may be considered as one of the most important steps in audio characterization and is similar to most pattern recognition problems. The basic audio features for a few sound samples stored in WAV files are computed and by using a measure of the reparability of the class by the use of histograms of the extracted features. [26-27]
The process of finding the features is as follows which extracts a particular feature and its statistics:
Features Statistics
Energy Entropy Standard Deviation (std)
Signal Energy Std by Mean (average) Ratio
Zero Crossing Rate Std
Spectral Rolloff Std
Spectral Centroid Std
Spectral Flux Std by Mean Ratio [34] [28-29]
4.5.1 Mel Frequency Cepstral Coefficient (MFCC) tutorial
The process of automatic speech recognition begins with computation of the unique features of the audio signal by which only the linguistic aspect of the sample can be identified. Hence by determining the shape of the vocal tract and by use of MFCCs we can correctly and accurately represent this envelope. [30-31]Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. We will deal with the main aspects of MFCCs, and how implement them. [32]
4.5.2. Feature extraction using multi signal wavelet Packet decomposition
It’s simply a feature extraction code using the wavelet packet Transform. [34] The code represents a generalization of the Multisignal 1-D wavelet decomposition.
4.5.3. Short time fourier transformation STFT and Inverse Short time fourier transformation ISTFT
The present code is a Matlab function that provides STFT of a given signal x(n) provides the following: stft - a matrix with complex stft coefficients with time across columns and frequency across rows, f - frequency vector, t - time vector.

5.0 PROPOSED RECONSTRUCTION METHODOLOGY

While implementing our system we may consider the following models which are already modified and are listed below:

5.1 Linear regression techniques on a z-transform implementation.

This idea consists of two voices, one is the source voice and other one is target voice. The first sample ie the target voice is the one in which form we try to observe the required input. The source voice is the sample which contains the information that we need to have reconstructed. [5] This method’s implementation consists of three major stages filter analysis, voice de filtering and voice conversion. The broad outline of each of these methods is as follows. In the first stage, we second using Machine Learning techniques such as minimizing the mean squared error, the components unique to any voice of human, subsequently this is what we refer to as the human voice filter. In the second stage, we use speech signal processing techniques like Z-transform to get the segment of the speech from the given speech signal,[33] by defiltering the unique voice content of the particular human voice. In the third and final stage, we now pass this de filtered voice into the

human voice filter of the target voice, and obtain the final speech in the target voice. [5]

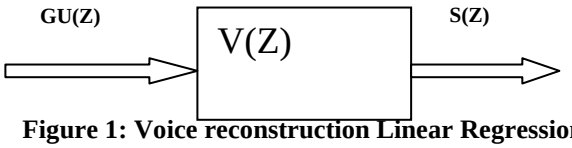


Figure 1: Voice reconstruction Linear Regression Techniques

5.2 The next basic idea is that of auto-regression on stationary time-frames where auto-regression is customized to the properties of the time-frame we consider. This is explained below. At stage 1 we implement the Dynamic Time Warping the second stage concentrates on K-Means clustering which emphasizes the fact that sound samples are stationary for relatively small time frames. This is justified by the fact that for small time frames, which are generally of 10ms, the sound varies very less. Each of the frames would have auto-regressive techniques performed on them.[6]

Stage three is Auto-Regression for time-frame where we use the auto-regression means on relatively stationary frames. While auto-regression process proposes that output samples are dependent on a few previous output and input samples. This uses a feedback from output to determine the future output samples. [7]

Stage four and five consists of Training phase and testing phase respectively where in the training phase we will perform clustering and will obtain coefficients where as in the testing phase we will start once the source speaker's voice sample is obtained. Then we will first split it into stationary time samples, as in the training phase. These stationary time samples, initially in our testing phase, are then detected to be part of a cluster, among the set of clusters obtained in the training phase. The output in the next stage is achieved by use of the cluster obtained. The second stage consists of predicting the output frame given the cluster the input frame. Once the cluster has been obtained, we pull out the coefficients corresponding to the cluster, and use it to linearly generate the samples which mimic the output.[6]

5.3 Cepstral Analysis

Before we start with the cepstral analysis part we can have a brief overview of the Speech Signal Analysis and this may be broadly classified into two basic steps. The first one is the fundamental frequency estimation in frequency domain and the second one is in the time domain.

5.4 The fundamental frequency estimation in frequency domain:

The basic problem associated with fundamental frequency determination is to consider a portion of the input signal and to trace the redundant dominant frequency. While doing so the problems that are encountered are that not all signals are repetitive and those of which are so may not be consistent in the time frame in which we are interested in. Next the signals may be associated with noise and a very interesting problem is that the repetitive signals with time interval of T are also periodic with interval 2T, 3T and hence forth hence we aim at finding the smallest repetitive sequence or the highest fundamental frequency and even signals of constant fundamental frequency may be changing in other ways over the interval of interest. The most trusted way of obtaining the unique fundamental frequency over the desired time frame for steady, noise free, and static speech signals is to use the cepstrum. [11]

5.5 Fundamental frequency estimation in time domain:

In this case the cepstrum is having a periodicity of log spectrum of the signal, but we are more interested in periodicity of the waveform itself. Here to get the fundamental frequency we take help of the autocorrelation. Expecting a proper correlation with itself for short delays is the crux. [11-12]

6.0 FORMANT FREQUENCY ESTIMATION

6.1 Prosody: The rhetorical flow of any word or its segment will definitely consider the correct intonation, proper stress at the punctuations and duration from written text is probably the most challenging problem for years to come. Estimation of formant frequencies is generally more difficult than estimation of fundamental frequency. The spectral shape of the vocal tract excitation strongly influences the observed spectral envelope, such that we cannot guarantee that all vocal tract resonances will cause peaks in the observed spectral envelope, nor that all peaks in the spectral envelope are caused by vocal tract resonances.[12]

The dominant method of formant frequency estimation is based on modelling the speech signal as if it were generated by a particular kind of source and filter:

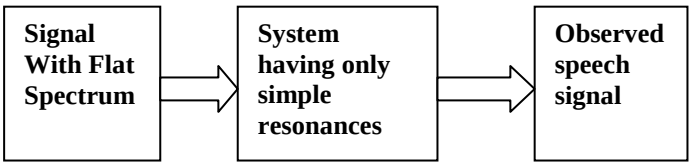


Figure 2: Estimation of formant frequencies

A generic human voice is basically a combination of excitation source and the vocal tract components or the system components. Now while going for the deconvolution process or the cepstral analysis our basic aim is to segregate the various speech components. [13] The convolution of the excitation and the unique traits of the vocal tract filter is the output speech signal.

If the excitation sequence $e(n)$ and the vocal tract filter sequence $h(n)$ then $s(n)$the speech sequence is:

$s(n) = e(n) * h(n)$

(1) The above mentioned can be mentioned in frequency domain as,

$s(\omega) = E\ (\omega) * H(\omega)$

6.2 The output is same in the time domain as expressed in Eqn 2 above. The system components and the multiplied source is transformed using cepstral analysis. [1]

6.2.1Cepstral analysis and its principle steps.
The provided speech spectrum's magnitude is as,

$|S(\omega)| = |E\ (\omega)| * |H(\omega)|$
6.2.3 The logarithmic representation of E(ω) & H(ω) will be,
$\log|S(\omega)| = \log|E\ (\omega)| * \log|H(\omega)|$
6.2.4 In Eqn. (4), for forther processing so as to take the summation the speech spectrum's log the excitation of the vocal tract component is taken into account. [16-17] While considering the vocal tract spectrum and the excitation spectra and its linear combination the bifurcation is performed by estimating the IDFT. Even though it changes but it remains same as that of the time domain.
$c\ (n) = \text{IDFT}(\log\ |S(\omega)|) = IDFT(\log\ |E\ (\omega)| + \log|H(\omega)|)$5. The basic cepstral analysis and the cepstral domain representation is shown below in Fig 3 and 4 respectively [14-15]. The results thus for the proper computation of the cepstrum is described in Fig 4, is given in Fig 5. In Fig 5, $s(n)$ is the voiced frame considered and $X(n)$ is the windowed frame. A multiplication by a hamming window to get $X(n) * |X(\omega)|$ in Fig 5 represent the spectrum of the windowed sequence $X(n)$. As the spectrum of the given frame is symmetric hence one half of the components is plotted. The $\log|X(\omega)|$ represents the log magnitude spectrum obtained by taking logarithm of the $|X(\omega)| * C\ (n)$ of Fig 5.
The various stages of cepstrum computation for an unvoiced frame is plotted in Fig 6.

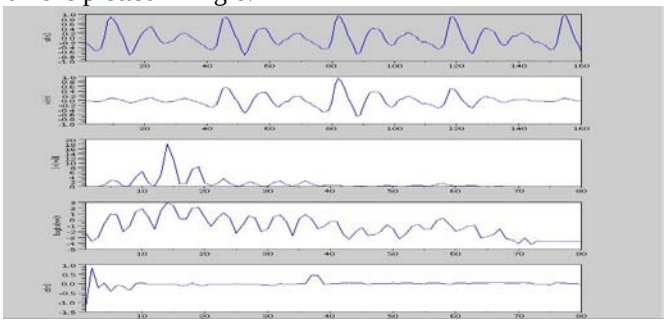


Figure 3: milli second cepstrum for voiced speech segment

6.2.5 Filtering operation or Liftering
In frequency domain when we perform filtering, we may call this overall process as Liftering. In Liftering we multiply the overall cepstrum with a rectangular window and this leads to our expected quefrency region of analysis at a desired time frame. High-time liftering and low time liftering are the two varients of liftering.

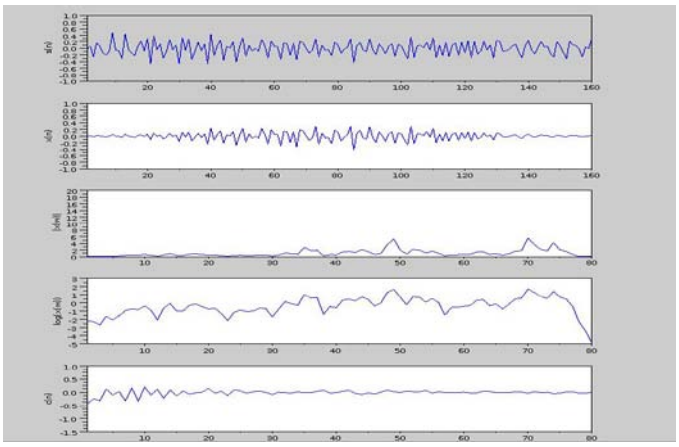


Figure 4: 20 milli second cepstrum for unvoiced speech segment

Low-time Liftering for Formant estimation
For quick varying vocal tract characteristics it's pretty easier to estimate the values but when the changes are relatively slower then low time liftering is used on the given speech sequence. This can be represented as follows,

$$W_e[n] = \begin{cases} 1,0 \le n \le L_c \\ 0, L_c \le n \le \frac{N}{2} \end{cases}$$

6.2.6. Where, L_c is the lower exclusion value of the liftering window,

$\frac{N}{2}$ is equal to the cepstrum length's half. In our case for the test case the value of L_c is 15 or sometimes 20. The parameters for the vocal tract characteristics are computed by the multiplication of cepstrum $C\ (n)$ and low time liftering window as mentioned in Eqn. (7) below.
$C_e(n) = W_e\ [n] * C(n)$
6.2.7. To get the log magnitude spectrum we have to have DFT applied on low time lifter. The output is basically the vocal tract spectrum of the provided short term speech and is denoted as below in Eqn. (8).
$\log[|H\ (W)|] = DFT\ [C_e(n)]$
The values of formant location and bandwidth can be calculated from vocal tract cepstrum. The heighest points of the smooth vocal tract spectrum are basically the values for the formant locations. The block diagram of formant estimation using low-time liftering is in Fig 7. Cepstrum of a voiced segment and low-time liftering window is shown in Fig 8, where as formant locations from vocal tract spectrum is depicted in Fig 9.

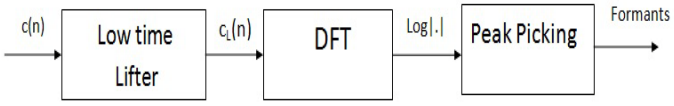


Figure 5: Block diagram representing low-time liftering

*6.2.*8. The values of formant location and bandwidth can be calculated from vocal tract cepstrum. The heighest points of the

smooth vocal tract spectrum are basically the values for the formant locations. The block diagram of formant estimation using low-time liftering is in Fig 4 and 5.

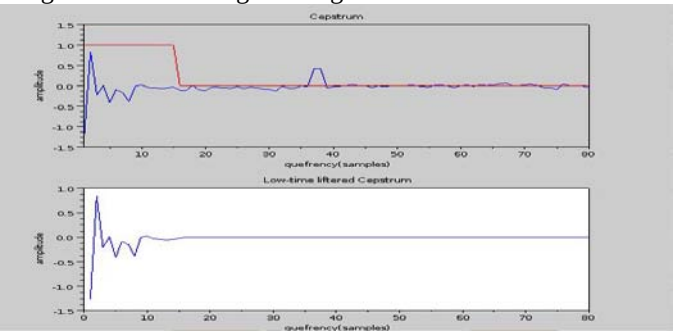


Figure 6: The voice segments Cepstrum

6.2.9. Liftering

The cepstrum liftering is the only way to find the vocal tract component and excitation component. The low-time liftering is applicable for the vocaltract component low-time liftering is done and high-time liftering is used for excitation component.

6.2.10. Computation of Formant frequency estimation:

This type of analysis is called source-filter separation, and in the case of formant frequency estimation, we are interested only in the modelled system and the frequencies of its resonances. To find the best matching system we use a method of analysis called Linear Prediction. Linear prediction models the signal as if it were generated by a signal of minimum energy being passed through a purely-recursive IIR filter. We will demonstrate the idea by using LPC to find the best IIR filter from a section of speech signal and then plotting the filter's frequency response.

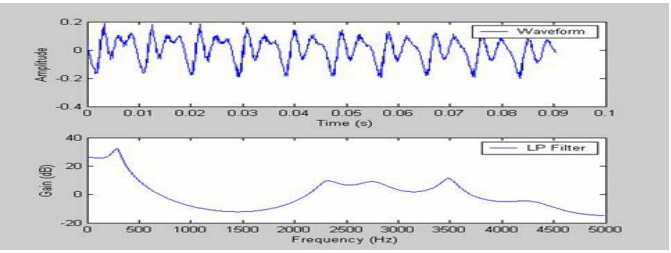


Figure 7: Formant frequency estimation

7.0 RESULTS

While plotting the fundamental frequency, the unique fundamental frequencies of speech are ovserved for their highest points in the corresponding quefrency region. In our case we searched for the the peak between 1 and 20ms in the whole cepstrum, which is clearly. The autocorrelation function for the speech signal's unique part can be seen in above incase of time domain estimation. The autocorrelation function is the peak when there is no latency and subsequently when the delays are ± 1 period, ± 2 periods, etc, we can estimate the it by finding out the highest point in the delay interval corresponding to the normal pitch range in speech, as in our case it is 2ms(=500Hz) and 20ms (=50Hz).

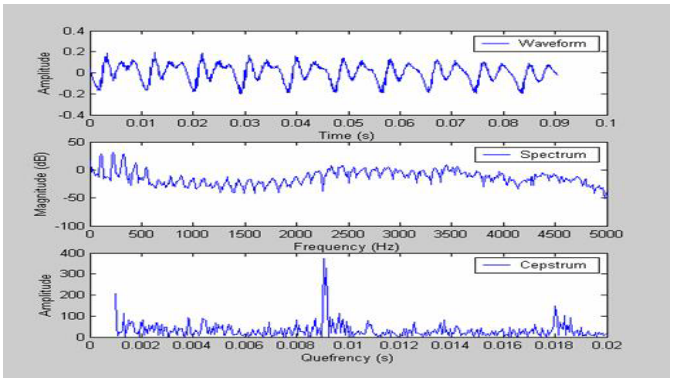


Figure 8: Fundamental Frequency Estimations

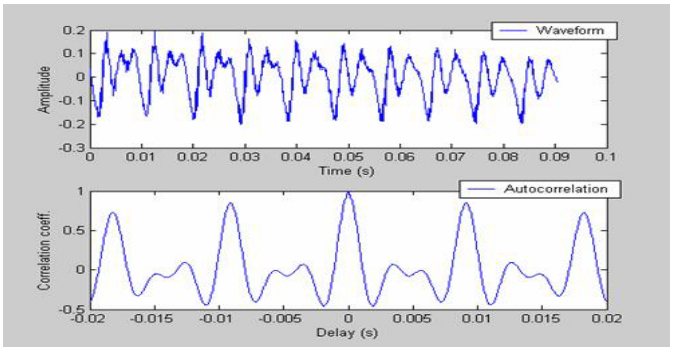


Figure 9: Fundamental Frequency Estimation in Time Domain

The cepstrum shows better results when the fundamental frequency is at the intermediate range and the fluctuations are not so high and also when the noise level is less. The cepstrum analysis is disadvantageous because of its high computation cost due to frequence domain processing. The autocorrelation function for the speech signal's unique part can be seen in Fig 8 & 9 incase of time domain estimation. The autocorrelation function is at the highest point when there is no delay and subsequently when the delays are ± 1 period, ± 2 periods, etc, we can estimate the it by finding out the highest point in the delay interval corresponding to the normal pitch range in speech, as in our case it is 2ms(=500Hz) and 20ms (=50Hz).

The autocorrelation approach works best when the signal is of low, regular pitch and when the spectral content of the signal is not changing too rapidly. The autocorrelation method is prone to pitch halving errors where a delay of two pitch periods is chosen by mistake. It can also be influenced by periodicity in the signal caused by formant resonances, particularly for female voices where F1 can be lower in frequency than Fx.

8.0 CONCLUSION AND FUTURE WORK

Even though in our present technological scenario we can observe a lot of applications of voice of speech processing in the form of recognizers, authentication systems, conversion

between speech to text and vice versa but the broad line benefits of voice processing may be a bit more exhaustive.
To have a better overview of these we may consider NLP as an area where we may use a syntactic parser on the input text and speech recognition's output may be used for information extraction techniques.

9.0 REFERENCES

[1]. Cepstral Analysis of Speech (Theory) : Speech Signal Processing Laboratory : Electronics & Communications : IIT GUWAHATI Virtual Lab.

[2]. P. S. Banerjee, Uttam Kumar Roy, "Modified PSOLA-Genetic Algorithm based approach for Voice Re-Construction", Journal on Information Technology, September – November 2013.

[3]. Voice Conversion for Unknown Speakers, Hui Ye and Steve Young Wiley India 2012.

[4]. Quality-enhanced Voice Morphing using Maximum Likelihood Transformations, Hui Ye, Student Member, IEEE, and Steve Young, Member, IEEE.

[5]. Reconstruction of human voice for impersonation Final report Amritha Raghunath Gunaa Arumugam Veerapandian Vignesh Ganapathi Subramanian 18 November, 2013.

[6]. Ye. H. and S. Young (2003).” Perceptually Weighted Linear Transformations for Voice Conver- sion”. Eurospeech 2003, Geneva.

[7]. Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E. & Russell, S. (In press). The Caterpillar: A Novel Reading Passage for Assessment of Motor Speech Disorders. American Journal of Speech Language Pathology.

[8]. Ganvit,Y Lokhandwala, MA and Bhatt, NS (2012). ”Implementation and Overall Performance Evaluation of Voice Morphing based on PSOLA Algorithm”, International Journal of Advanced Engineering Technology.

[9]. “Pitch Conversion Based on Pitch Mark Mapping” Srikanth Mangayyagari and Ravi Sankar Department of Electrical Engineering, University of South Florida, Tampa, FL 33620, USA E-mail: {smangayy, sankar}@eng.usf.edu

[10]. Patel, R., Hustad, K. Connaghan, K.P. & Furr, W. Relationship Between Prosody and Intelligibility in Children with Dysarthria. Journal of Medical Speech Language Pathology.

[11]. http://svr-www.eng.cam.ac.uk/~ajr/SA95/node34.html

[12]. http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html

[13]. R. Deller Jr., J.H.L. Hansen, J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, 2000.

[14]. A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-Time Signal Processing, U. Saddle River, Ed. NJ:Prentice Hall, 1999.

[15]. L.R. Rabiner and R.W. Schafer, Theory and Application of Digital Speech Processing, First

[16]. Edition, Prentice Hall, New York, 2011.

[17]. D. O`Shaughnessy, Speech Communications: Human and Machine, Second Edition, University Press, India, 2004.

[18]. L.Rabiner, B.Juang, B.Yegnanarayana, Fundamentals of speech recognition, Pearson, India, 2010.

[19]. Urmila Shrawankar, Dr. Vilas Thakare “Techniques For Feature Extraction In Speech Recognition System : A Comparative Study” M.Tech dissertation Amravati University, 2011.

[20]. J. W. Picone, "Signal modelling technique in speech recognition," Proc. Of the IEEE, vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[21]. Shanthi Therese S.,Chelpa, “Lingam Review of Feature Extraction Techniques in Automatic Speech Recognition” International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.2, Issue No.6, pp : 479-484 1 June 2013.

[22]. Lawrence R. Rabiner, et. Al. Speech Recognition by Machine. 2000 CRC Press LLC.

[23]. Meysam Mohamad pour, Fardad Farokhi, “An Advanced Method for Speech Recognition”, World Academy of Science, Engineering and Technology 25, 2009.

[24]. Simon Kinga and Joe Frankel, Recognition, “Speech production knowledge in automatic speech recognition”, Journal of Acoustic Society of America, Oct 2006.

[25]. Tickle A, Raghu S and Elshaw M, “Emotional recognition from the speech signal for a virtual education agent” Sensors & their Applications XVII, IOP Publishing, Journal of Physics: Conference Series 450 (2013).

[26]. Eyden, F., Wollmer, M., and Schuller, B. 2010 Opensmile: the munich versatile and fast opensource audio feature extractor, MM '10 Proceedings of the international conference on Multimedia, pp. 1459-1462.

[27]. R. N. Khushaba, A. Al-Jumaily, and A. Al-Ani, “Novel Feature Extraction Method based on Fuzzy Entropy and Wavelet Packet Transform for Myoelectric Control”, 7th International Symposium on Communications and Information Technologies ISCIT2007, Sydney, Australia, pp. 352 – 357.

[28]. R. N. Khushaba, S. Kodagoa, S. Lal, and G. Dissanayake, “Driver Drowsiness Classification Using Fuzzy Wavelet Packet Based Feature Extraction Algorithm”, IEEE Transaction on Biomedical Engineering, vol. 58, no. 1, pp. 121-131, 2011.

[29]. J. Benesty, M. Sondhi, Y. Huang. Springer Handbook of Speech Processing. Berlin, Springer, 2008.

[30]. B. Boashash. Time Frequency Signal Analysis and Processing: A Comprehensive Reference. Oxford, Elsevier, 2003.

[31]. T. Dutoit, F. Marqures. Applied Signal Processing: A MATLAB-Based Proof of Concept. New York, Springer, 2009.
[32]. J. Allen. "Application of the short-time Fourier transform to speech processing and spectral Analysis". Proc. IEEE ICASSP-82, pp. 1012-1015, 1982.
[33]. J. Smith III, X. Serra. "PARSHL:An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation". Tokyo, Proceedings of the International Computer Music Conference (ICMC-87), pp. 290 – 297, 1987.
[34]. Vogt, T., André, E., and Bee., N. 2008 EmoVoice - A framework for online recognition of emotions from voice. In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee.
[35]. www.mathwork.com
[36]. arxiv.org
[37]. dspace.thapar.edu:8080
[38]. Cepstral Analysis of Speech (Theory): Speech Signal Processing Laboratory: Electronics & Communications: IIT GUWAHATI Virtual Lab.

E-Commerce and Economy: A Case Study of Saudi Arabia

MotebAyeshAlbugami[1]

Submitted in April, 2015; Accepted in May, 2015

Abstract –Evolution of E-commerce has revolutionized the business world. As in many great discoveries and innovations, initially there was a lot of skepticism about the concept and the details of its implementation. It took a while for the masses to embrace the concept and start buying online. Of course, E-Commerce is a byproduct of internet, which became available to the public in 1991.Immidiately after the availability of internet to public; the concept of e-Commerce was mooted which was embraced by some businesses. In the beginning, the response internet usage was slow and that too was the case with e-Commerce as well. Today the number of internet users exceeds three billon and those of the e-Commerce buyers are around one and a quarter billion. Economic impact of e-Commerce on global and national economies has been tremendous. According to market, research, projected revenue from e-Commerce sales in 2015 is expected top 24 trillion dollars. About two thirds of the internet users come from developed world; and so, it is natural to expect that the bulk of e-Commerce buyers also come from developed world. The aim of this research is to study the extent of E-Commerce in Saudi Arabia and its impact to the Saudi economy. In the course of this study, we shall outline the underlying factors in the development and progress of E-commerce in Saudi Arabia.

Index Terms – E-Commerce, Web 2.0, Saudi Arabia Saudi Economy, Saudi Business

1.0 INTRODUCTION

Commercial E-Commerce started soon after the introduction of internet in 1991. Emergence of internet has revolutionized many facets of life. In particular, internet has changed the ways of transacting business,removing many formalities and setting new rules of trade. E-Commerce has transformed the world into a global business village. In the process it has and continues to have a profound impact on the world and regional economies. E-Commerce revolution continues as it embraces new technologies and extends itself to more countries and regions. As expected, the affluent societies, who could afford internet at its early availability are excelling in and benefiting from the E-Commerce. This is evident from the available data. Saudi Arabia, a major exporter of oil, is one of the G20 countries. Saudi Arabia is also a regional power in the Middle East and the Arab world. The aim of this study is to explore the extent of E-Commerce activity in Saudi Arabia and its impact on Saudi Economy.

[1]*Department of Management Information Systems, Faculty of Business Administration, Tabuk University, Saudi Arabia*
E-mail: Malbugami@ut.edu.sa

2.0 AN OVERVIEW OF E-COMMERCE

E-Commerce is a bye product of Internet. Up until the eighties of 20th century we hadn't imagined buying items of our daily needs without viewing and checking them physically. It was unimaginable to think of buying commodities like flowers and fragrances without smelling and looking at them physically. With the invention of internet, we now indulge in purchasing almost everything online, not only is what available locally but also from the national and international markets, demolishing the age old boundaries and barricades. Like in many other cases, e-commerce has had its journey to the current level of acceptance as a way of shopping. With a hesitant start in the early nineties of the last century, the e-commerce has now become a way of life in the developed countries. Even in the developing countries, more and more people are embracing the travel and hassle free way of shopping. More importantly, the E-commerce has defined a new paradigm for global trade, making it possible for individuals to make purchases globally and receive their purchased items as shipments from almost anywhere in the world. The number of e-commerce participants, both the traders and buyers is steadily increasing. Many innovations are being explored to make e-commerce even more attractive. Indeed, buying online reduces, in some cases drastically, prices of many items. E-commerce is now a well-developed field of study and is taught in most of the educational institutes around the world. Most of the details can be found in any standard text on the same topic, for example [1].

2.1 E-Commerce Statistics

According to [**2**], the number of internet users worldwide in May 2015 was over 3.1 billion. About 66% of these users were from developed countries and only 34% are from developing countries. From the Unites States of America (USA) alone, the number of internet users was about 2.1 billion, accounting for 75% of the total customer market. With the growth of Web 2.0, the number of e-Commerce first grew exponentially and now has sustained a steady increase. According to[3], the total number of E-Commerce buyers in the world in 2015 is estimated to be 1.2 billion. According to [4], the number of digital buyers in the USA in 2014 was over 140 million. This is just about 12% of the total number of e-commerce customers worldwide, suggesting that the USA has a fairly low percentage of the e-commerce buyers worldwide which is somewhat surprising. The total number of buyers in the world and the USA, and their future projection are shown in Fig 1and Fig 2.

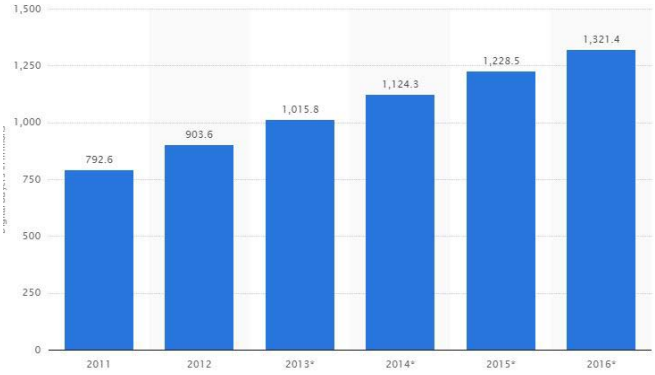


Figure 1: Number of E-Commerce buyers globally

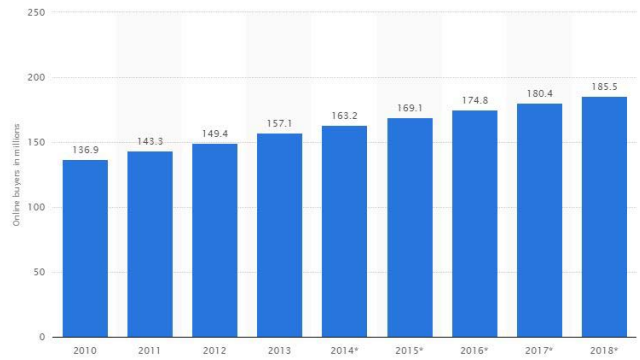


Figure 2: Number of E-Commerce buyers in USA

The growth of e-commerce has been exponential since its early days in 1995. The projections for retail sales, as seen in fig.3, are somewhat staggering. It is expected that the retail Sales worldwide will top to twenty two Trillion this Year [5]. Another source [6] puts this figure to 25 trillion.

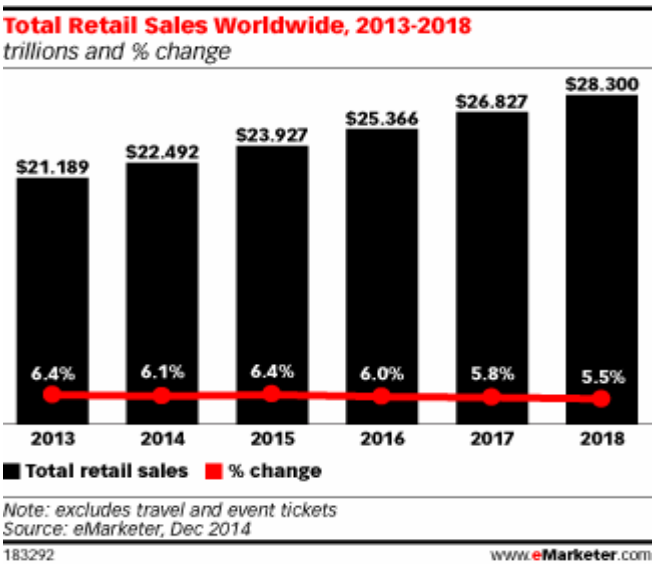


Figure 3

Since its commercial availability, the internet has provided huge opportunities and benefits to the society and businesses alike by simplifying its accessibility and processes. From e-commerce point of view, the world has now become a virtual village. Nations, regions and societies around the world are benefiting from globalization of the e-commerce. More than half of the world population is yet to benefit from the revolution of internet and its fruits like e-commerce, e-business, e-learning etc. The reasons for the inability of these societies to acquire connections and indulge in e-commerce are political, economic, educational and social. For example, if a government of particular country doesn't provide infrastructure and delays enacting laws, the people themselves can do nothing. Other hurdles in the way of making e-Commerce within a common man's reach submerged in the poverty. Many countries in Asia and Africa are still struggling for basic needs of livelihood like food, drinking water and healthcare. For these societies e-Commerce is still a distant dream. Due to their economic condition, these societies and their governments are bound to take a much longer time to make e-Commerce a common man's choice.

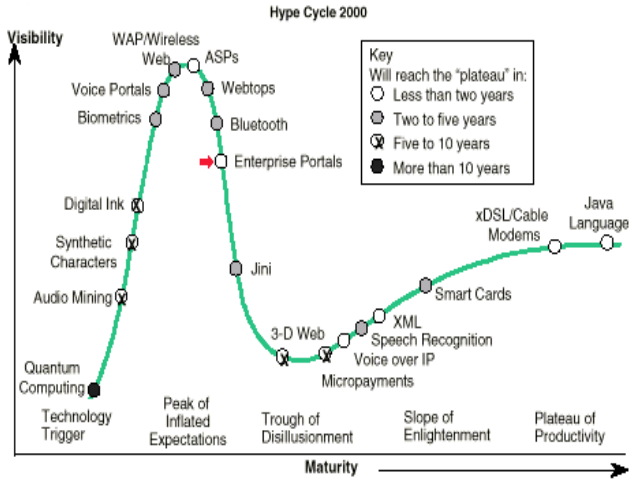


Figure 4: Technology Hype Cycle 2000

Gartner Hype Cycles [7] have been monitoring the pulse of technologies globally. In particular, Gartner Hype Cycles provide emerging technologies and their trends. Hype Cycles help enterprises evaluate the suitability of various e-commerce technology capabilities hyped in the market and understand their maturity and business value.Figures 4-9 [7], depict e-commerce in the midst of technological advancement. In the year 2000, e-commerce as seen in Fig 4, when still in infancy, e-commerce emerged as a one of the technology or service of the overall technological developments of the time. Since availability of internet and associated technologies are the main sources and media for the provision of E-Commerce, gradual emergence of these technologies is noticeable in the Hype Cycles. As we can see in Fig 5, it was, much later, in year 2010, when the first dedicated Hype Cycle emerged solely for e-commerce. Since then Garner have regularly provided E-Commerce Hype Cycles as can be seen in Fig 6-9. In other words the e-commerce has now become a well-establish and

accepted way of shopping of small and large items alike. As the demand for e-commerce grows, we can expect more innovations and improvements in the technologies associated with e-commerce. .

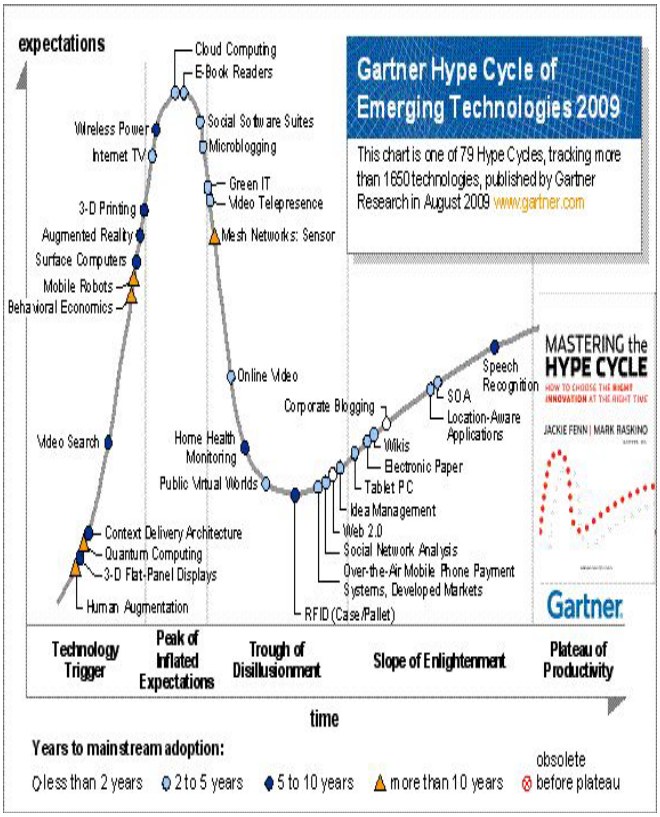


Figure 5 E-Emerging Technologies Hype Cycle 2009

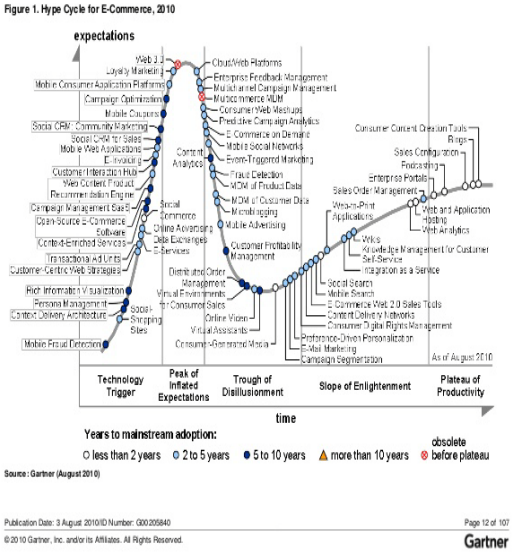


Figure 6: E-Commerce Hype Cycle 2010

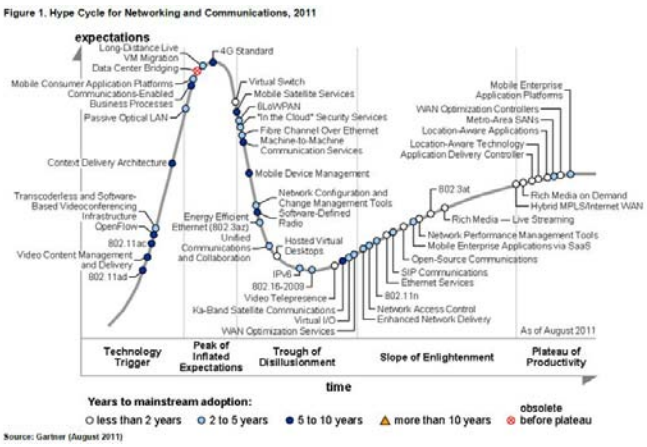


Figure 7: E-Commerce Hype Cycle 2011

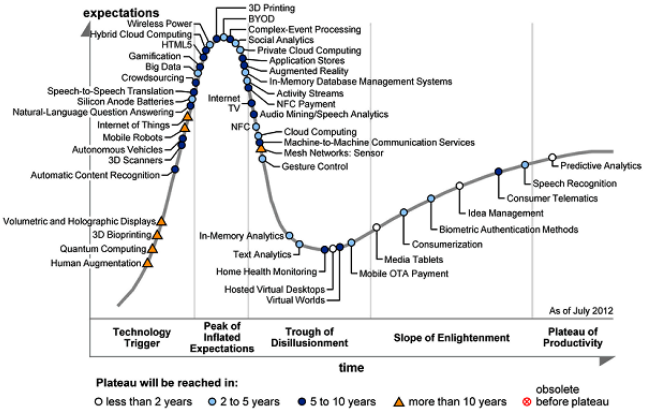


Figure 8: E-Commerce Hype Cycle 2012

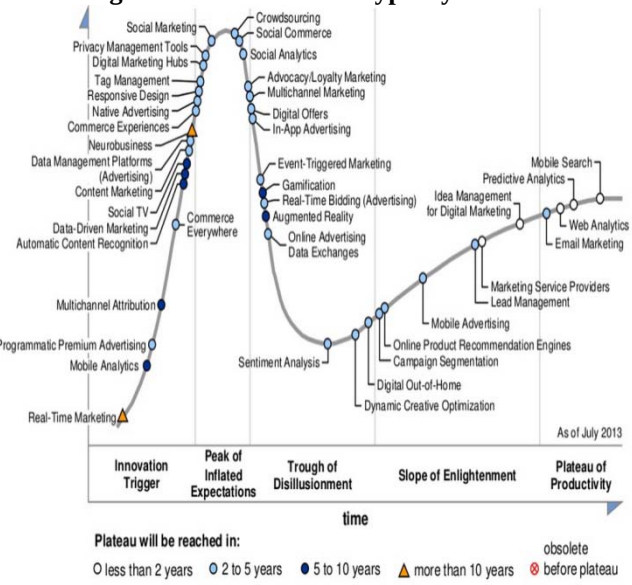


Figure 9: E-Commerce Hype Cycle 2013

With the emergence of iPhone in 2007 and an explosion in the availability of other similar devices, the concept of m-commerce has been revolutionised digital commerce. Now m-commerce is growing at an exponential rate. Its cut into the traditional e-commerce market, taking more and more hare of it

into m-commerce. As seen in Fig 10, taken from [8], m-commerce.

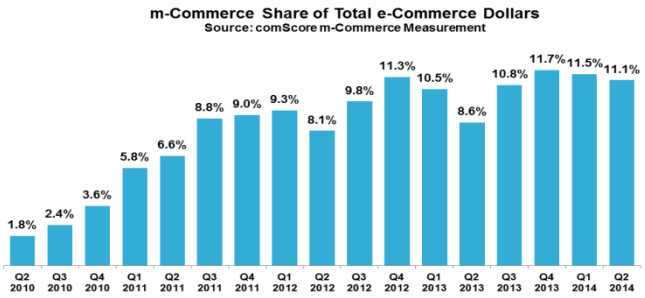


Figure 10: Percentage of m-commerce in e-commerce

commerce is becoming more and more popular. As the mobile technology becomes more refined, m-commerce is expected to swallow a much larger share of the e-commerce and traditional way of commerce.

3.0 SAUDI ARABIA: ECONOMICAL AND SOCIAL BACKGROUND

Saudi Arabia is a major producer and exporter fossil fuel and is a member of G20. In terms of industry, especially small and medium scale industries, it is regarded as a developing country.

Questions	Average	Questions	Average
Question 1	5	Question 8	4
Question 2	4	Question 9	4
Question 3	4	Question 10	4
Question 4	4	Question 11	4.3
Question 5	4	Question 12	4.1
Question 6	4	Question 13	4.7
Question 7	4		

For a detailed discussion of Saudi industrial activity, please see [9] and [10]. Per capita GDP of Saudi Arabia is shown in Fig 11, taken from [11].

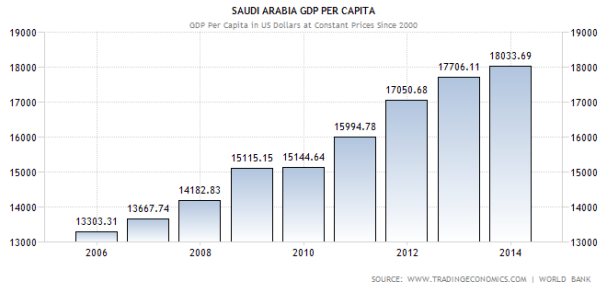


Figure11: Per capita GDP of Saudi Arabia

According to [13], "Saudi Arabia is the largest economy in the entire Arab world. It has the largest IT market in the region and has a very high youth It has biggest retail market in the Middle East. Saudi Arabian population's usage of internet is growing exponentially as can be seen in Fig 12. According to [14], Saudi Arabia has registered the MENA's (Middle East and North Africa) highest e-commerce growth rate in Q1 2014, according to the latest analysis by Visa. The kingdom saw an estimated 43 percent overall e-commerce growth, comparing 2013 and 2014 data, the retail electronic payments network's study revealed. The Q1 2014 growth was driven by increases in both domestic and cross-border e-commerce, which saw a 67 percent and 36 percent growth over the same period last year, respectively. Emerging as the leading categories for spending were general department store and airline transactions, followed by travel agencies, financial services and fashion retail." The kingdom has been one of the leading markets in the GCC to have embraced electronic payments alongside rapid internet and broadband penetration, which has resulted in greater adoption of financial cards for e-commerce transactions [14].

There are social and religious factors which play significant part in the business and commerce activity of Saudi Arabia. Makkah in Saudi Arabia houses sacred Kaaba, which is visited by an estimated 15 million people in a year, which generates a significant amount of revenue [15]. The well-known activity of hajj takes place every year in Saudi Arabia, which is regarded as the seat of Islam in the same manner as Vatican.

4.0 OUR SURVEY

In order to gauge the e-commerce activity and participation of SMEs in Saudi Arabia, we have conducted a survey on the West coast of Saudi Arabia, mainly in the port city of Jeddah, which regarded to be a centre of Saudi Trade and commerce. The aim of our survey was to find out involvement of SMEs in e-commerce acquisition and provision. The following are the questions, which were provided response by 97 participating sanitary ware SMEs.

1. My company doesn't have an Accounting & Sales system and that makes hard to manage the business
2. I have an Accounting & Sales system and it is suitable for my company needs
3. An Accounting & Sales system is required for smooth functioning of my company
4. My Accounting & Sales comes with functions that exceed my company needs.
5. Working with my Accounting & Sales system requires little or no training to use it
6. There are privacy issues with my Accounting & Sales in relation to customer data
7. There are privacy issues with my Accounting & Sales in relation to enterprise data
8. There are data security issues associated with my Accounting & Sales system
9. My Accounting & Sales system is missing some important functionalities
10. My enterprise lacks qualified staff to use an Accounting & Sales system
11. My enterprise has adequate IT facilities to deal with the requirements of my Accounting & Sales system

12. Frequency of technical support required by my Accounting & Sales exceeds my expectations
13. I am satisfied with the performance of my system

Survey used a scale of five choices: 1 for highly disagree, 2 for disagree, 3 for neutral, 4 for agree and 5 for highly agre. After carefully analyzing the survey responses of ninety seven companies, the averages of their responses are provided in the following Table 3.

Analysis of the surveys

From the results of the survey it is evident that a vast majority of sanitary ware industries in Jeddah and the surrounding areas do not use management and accounting systems. Many of the managers of these companies have admitted to the fact that if they were using MIS systems, they would be indulging in e-commerce. If this is taken as a guide, we can say that small hardware industry is far from being close to embracing and providing e-commerce platforms. The Saudi Arabian government has been providing all what is required for businesses to enter into the e-commerce [16]. Therefore, there are no excuses for not using or implementing IT systems and tools which would enable e-commerce activity in these SMEs.

Table 3: Survey Results

Questions	Average	Questions	Average
Question 1	5	Question 8	4
Question 2	4	Question 9	4
Question 3	4	Question 10	4
Question 4	4	Question 11	4.3
Question 5	4	Question 12	4.1
Question 6	4	Question 13	4.7
Question 7	4		

5.0 CONCLUSION

From the discussion and survey results, it is evident that Saudi Arabia has a significant amount of e-commerce activity, and is a leader in the region, GCC, the Middle East and the Arab world. The most of e-commerce activity appears to be in big companies in the sector of communication, civil aviation and large industries. However, there is negligible amount of participation by SMEs in the e-commerce. There is ample evidence that the government of Saudi Arabia has been proactive in making internet available to most of its people even in the remote regions. It is high time for the SMEs to realise the power and benefits of e-commerce and hence pave the way to participate in this model of business.

REFERENCES

[1]. Kenneth C. Laudon, E-Commerce, Prentice Hall, ISBN-10: 0133507165, ISBN-13: 9780133507164 pp 912, 2014

[2]. Internet Live Stats, [Online], Available: http://www.internetlivestats.com/internet-users/ (April 1, 2015)

[3]. Statista – the statistical portal, [Online], Available from http://www.statista.com/statistics/251666/number-of digital-buyers-worldwide/, (April 1, 2015)

[4]. Statista – the statistical portal, [Online], Available, http://www.statista.com/statistics/273957/number-of-digital-buyers-in-the-united-states/, (April 1, 2015)

[5]. Marketer, Retail Sales Worldwide Will Top $22 Trillion This Year, [Online], Available fromhttp://www.emarketer.com/Article/Retail-Sales-Worldwide-Will-Top-22-Trillion-This-Year/1011765#sthash.99nC43bA.dpuf, (April 1, 2015)

[6]. My Customer, Global retail sales set to reach $24 trillion in 2015, [Online], Available from http://www.mycustomer.com/news/global-retail-sales-set-reach-24-trillion-2015, (April 1, 2015)

[7]. Gartner, Gartner Hyper Cycle, [Online], Available from http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp, (April 1, 2015)

[8]. ComeScore, Q2 M-Commerce Explodes to 47% Y/Y Gain: What it Means for the Growth of Mobile, [Online], Available from http://www.comscore.com/Insights/Blog/Q2-M-Commerce-Explodes-to-47-YY-Gain-What-it-Means-for-the-Growth-of-Mobile, (April 1, 2015)

[9]. Yamin, M. & Al Hudhaif, A. A. (2014). MIS in Small Industry: Sanitary Ware in Saudi Arabia. International Journal of Emerging Technology and Advanced Engineering, 4 (11), 487-494

[10]. Yamin, M. & Al Makrami, A. A. (2015). Cloud Computing in SMEs: Case of Saudi Arabia. BIJIT - BVICAM's International Journal of Information Technology, 7 (1), 853-860

[11]. Trading Economics, Saudi Arabia GDP per capita, [Online], Available from http://www.tradingeconomics.com/saudi-arabia/gdp-per-capita, (April 1, 2015)

[12]. Electronic Newsletter, 51 Million Mobile Subscriptions in Saudi, [Online], Available from Arabiahttp://www.citc.gov.sa/English/MediaCenter/Newsletter/Documents/PR_ENL_017.pdf, (April 1, 2015)

[13]. YUZR TECHNOLOGIES, The Scope of E-Commerce Business In Saudi Arabia, , [Online], Available from http://www.yuzr.com/the-scope-of-e-commerce-business-in-saudi-arabia/,(April 1, 2015),

[14]. Business.Com, 43% e-commerce growth in Saudi Arabia, [Online], Available from

http://www.arabianbusiness.com/43-e-commerce-growth-in-saudi-arabia-553766.html, (April 1, 2015)

[15]. Yamin, M. (2015). Health Management in Crowded Events: Hajj and Kumbh. BIJIT - BVICAM's International Journal of Information Technology, 7 (1), 791-794.

[16]. Ahmed Al Saleh , Exploring Strategies for Small and Medium Enterprises in Saudi Arabia, Strategies for SMEs in Saudi Arabia, [Online], Available: http://www.ribm.mmu.ac.uk/symposium2012/extendedabstracts/AhmedAlSaleh.pdf

BIJIT - BVICAM's International Journal of Information Technology

(A Half Yearly Publication; ISSN 0973 - 5658)

Subscription Rates (Revised w.e.f. January, 2012)

Category	1 Year		3 Years	
	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**

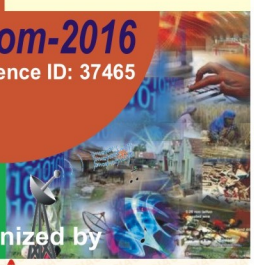
Place: (with official seal)

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

1964-2014
celebrating
Golden Jubilee
BHARATI VIDYAPEETH
Founder Hon'ble Dr. Patangrao Kadam

INDIACom-2016
IEEE Conference ID: 37465



Organized by

BHARATI VIDYAPEETH PUNE

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

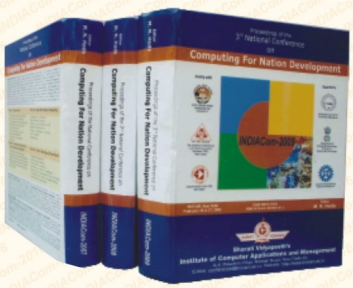
GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY

The Institution of Electronics and Telecommunication Engineers (IETE), Delhi Centre

The Institution of Engineering and Technology Delhi Local Networks

CSI Region - I & Divisions - I, II, III, IV & V

ISTE, Delhi Section



(Copies of the proceedings of past *INDIACom*s)

Correspondence

All correspondences related to the conference must be sent to the address:-

Prof. M. N. Hoda
General Chair, *INDIACom - 2016*
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, indiacom2016@gmail.com
visit us at: http://www.bvicam.ac.in/indiacom

INDIACom-2016

10th INDIACom;2016 3rd International Conference on

Computing for Sustainable Global Development

(16th-18th March, 2016)

IEEE Conference Record Number # 37465

Paper Submission Link : http://www.bvicam.ac.in/indiacom/ SubmitPaper.asp.

INDIACom-2016 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 use in enhancing the quality of life and contribute effectively to realize the nations' vision of sustainable inclusive development using Computing. It is an amalgamation of four different Tracks organized parallel to each other, in addition to few theme based Special Sessions, as listed below:-

Track #1: Sustainable Computing
Track #2: High Performance Computing
Track #3: High Speed Networking and Information Security
Track #4: Software Engineering and Emerging Technologies
Track #5: Theme Based Special Sessions

INDIACom-2016 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, special 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 Networks
- Mobile Computing
- Entertainment Technologies
- Multimedia Computing
- Information and Collaboration Systems
- Fuzzy, Soft Computing and Nature Inspired 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 original papers online at http://www.bvicam.ac.in/indiacom/ SubmitPaper.asp. Only electronic submissions will be considered. Papers submitted through E-mail, as attachment, will not be considered.

Review Process, Publication and Indexing

All the submitted papers shall be doubled blind reviewed, by 03 experts, on the basis of their technical suitability, scope of work, plagiarism, 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 978-93-80544-19-9** serials. Conference proceedings will also be available in soft copy having **ISBN 978-93-80544-20-5** serial. All accepted papers, which will be presented in the conference, will be submitted for publication and indexing to **IEEE Xplore.**

Important Dates

Submission of Full Length Paper	**10th November, 2015**	**Paper Acceptance Notification**	**12th January, 2016**
Submission of Camera Ready Copy (CRC) of the Paper	**25th January, 2016**	**Registration Deadline** (for inclusion of Paper in Proceedings)	**01st February, 2016**