Global Journal of Computer Science and Technology

discovering thoughts and inventing future

Volume 10 Issue 12 Version 1.0 Online ISSN: 0975-4172 Print ISSN: 0975-4350

14 Technology Reforming Ideas

highlights

Clustering And Semantic Similarity

Wireless Network

Three Dimensional Database

Neuro-Fuzzy Model

October 2010

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 10 Issue 12 (Ver. 1.0)

Global Association of Research

This is a special issue published in version 1.0 of "Global Journal of Computer Science and Technology." By Global Journals Inc.
All articles are open access articles distributed under "Global Journal of Computer Science and Technology"

Reading License, which permits restricted use. Entire contents are copyright by of "Global Journal of Computer Science and Technology" unless otherwise noted on specific articles.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without written permission.

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://www.globaljournals.org/global-journals-research-portal/guideline/terms-and-conditions/menu-id-260/.

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374 Import-Export Code: 1109007027

USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*
Sponsors: Global Association of Research
Open Scientific Standards

Publisher's Headquarters office

Global Journals Inc., Headquarters Corporate Office, Cambridge Office Center, II Canal Park, Floor No. 5th, ***Cambridge (Massachusetts)***, Pin: MA 02141
United States
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Inc., City Center Office, 25200 Carlos Bee Blvd. #495, Hayward Pin: CA 94542
United States

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org* Investor Inquiries: *investers@globaljournals.org* Technical Support: *technology@globaljournals.org* Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 500 USD (Color)

Editorial Board Members (HON.)

President Editor (HON.)

Chief Author (HON.)

Dean & Editor-in-Chief (HON.)

Contents of the Volume

From the Chief Author's Desk

We see a drastic momentum everywhere in all fields now a day. Which in turns, say a lot to everyone to excel with all possible way. The need of the hour is to pick the right key at the right time with all extras. Citing the computer versions, any automobile models, infrastructures, etc. It is not the result of any preplanning but the implementations of planning.

With these, we are constantly seeking to establish more formal links with researchers, scientists, engineers, specialists, technical experts, etc., associations, or other entities, particularly those who are active in the field of research, articles, research paper, etc. by inviting them to become affiliated with the Global Journals.

This Global Journal is like a banyan tree whose branches are many and each branch acts like a strong root itself.

Intentions are very clear to do best in all possible way with all care.

Dr. R. K. Dixit
Chief Author
chiefauthor@globaljournals.org

Using Product Similarity for Adding Busines Value and Returning Customers

Boumedyen A.N. Shannaq[1], Prof. Victor V. Alexandrov[2]

GJCST Classification
H.3.3, H.3.1

***Abstract*-Due to increasing attentionto maximize profits, international firms and corporations oversee the importance of typing and registering name of products. It occurs that the product name sometimes misspelled by customers during the product registration. The customer finds it difficult to search for a product on the internet either because the product is not registered or is not documented in the right order. This study highlights the problem and creates alternative ways to retrieve similar product name. To experiment this idea, a collection of English and Arabic product names have been built, along with 93 training queries and 123 test queries. These collected data are used to evaluate a variety of algorithms to measure effectiveness of using information retrieval operation. The new technique LIPNS shows a considerable improvement over existing.**

Keywords -Names Similarity, Arabic names, SoundX, N-gram, and Information Retrieval.

I. Introduction

Finding regularities in strings is useful in a wide area of applications which involve string manipulations. Such applications add richer data-profiling capabilities to its data quality offerings, to increasing its customer base in Europe [1]. The task of matching entity names has been explored by a number of communities, including statistics, databases, and artificial intelligence. Each community has formulated the problem differently, and different techniques have been proposed. Finding correspondences between elements of data schemas or data instances is required in many applications. This task is often referred to as matching. A comparison shopping website that aggregates product offers from multiple independent online stores. The comparison site developers need to match the product catalogs of each store against their combined catalog. Names and addresses are critical in identifying a person, a company, an organization etc. This information is the primary keys for accessing the information of an individual or a company in many of the database system that exist in the computer world. The variation of names, the variation in the way they are written or spelt creates major problem to name recognition across the globe. Product names have characteristics that make them different to general text. While there is only one correct spelling for many words, there are often several valid spelling variations for product names, for example _Tomato', and _Tamato'.Names are also heavily influenced by people's cultural backgrounds. These issues make matching of personal names more challenging compared to matching of general text [2].There are many applications of computer-based name-matching algorithms including record linkage and database searching where variations in spelling, caused for example by transcription errors, need to be allowed for. The success of such algorithms is measured by the degree to which they can overcome discrepancies in the spelling of names. Evidently, in some cases it is not easy to determine whether a name variation is a different spelling of the same name or a different name altogether. Most of these variations can be categorized as Spelling variations, Phonetic variations, Double names and Double first names [3].Now, International firms and corporations oversee the importance of different customers represent different levels of profit for the firm especially Gulf Arabic customer's, who don't know very well, how to type correct product name and they have to make their best guess at how to type the product names correctly. Because if they misspelled product name, exact match search will not find product in the DB, subsequently, the customer will not be willing to use this system again. At the same time, the number of customers will stop purchasing products or services from this company. It is an important indication of the growth or decline of a firm's customer base. Product name search in particular, however, to our knowledge has not been studied. In [4] they have discussed the characteristics of personal names and the potential sources of variations and errors in them, and presented an overview of both pattern matching and phonetically encoding based name matching techniques. There Experimental results on different real data sets have shown that there is no single best technique available. The characteristics of the name data to be matched, as well as computational requirements, have to be considered when selecting a name matching technique.However, we have built a collection of test English and Arabic product names and queries with corresponding relevance judgment; developed a new technique Language-Independent Product Name Search (LIPNS), discuss the results of a series of comparison experiments to see which matching Techniques achieve the best matching quality for different name types, and to compare their computational performance with LIPNS, all name matching techniques were implemented and compared with LIPNS. The obtained results show that new LIPNS technique provides an improvement over variant techniques.

About[1]- Information Systems DepartmentUniversity of Nizwa , Sultanate of Oman Email: boumedyen@unizwa.edu.om
Email: boumedians@yahoo.com
About[2]- Academy of Natural Sciences, Russia Doc.Nat.Sci., St.Petersburg Institute for SPIIRAS, St. Petersburg, Russia E-mail: alexandr@iias.spb.su

II. Statement of problem

Using product name to retrieve information makes these systems susceptible to problem arising from typographical errors. These exact match search approach will not find instances of misspelled product name or those product names that have more than one accepted spelling. The importance of such name-based search algorithms has resulted in improved name matching algorithms for English that make use of phonetic information. But these language-dependent techniques have not been extended to other languages such as Arabic, Chinese, and Indian etc. In the existing Soundex name search algorithm, the name search is limited to only English. But, the n-gram matching algorithm is not limited to any language and to our knowledge it is not applied for Arabic language name search. The limitation of this work in this area is partly due to the lack of standardized test data. Hence, we developed a collection of 17,265 Arabic and English product names and personal names, along with 93 training queries and 123 test queries. We use this collection to evaluate Soundex, n-gram, including new LIPNS method proposed to calculate the effectiveness of this standard information retrieval measures.

1) Selecting Algorithms

Numerous name search algorithms for Latin-based languages exist that effectively find relevant identification information, that use the phonetic features of names have been researched thoroughly for English, while string similarity techniques have garnered interest because of their language-independent methodology. Identify matching systems frequently employ name search algorithms to effectively locate relevant information about a given product name. In this study, we select the best suited algorithm for name search matching and the comparison is done with LIPNS technique . In [4] and [5] A comparison of various name matching algorithms through the R&W database is described in the Figure 2.1.

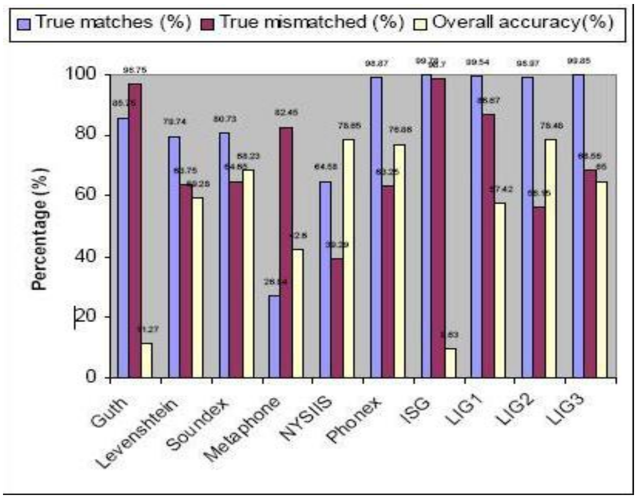


Figure 2.1 A comparison of various name matching algorithms [5].

As shown in Figure 2.1 above the algorithms accuracies are good fro Phonex, Soudex and LIG2.

Table 2.1 Average f-measure values (best results shown boldface and worst results underlined) [4].

	Given names
Soundex	.342
Phonex	.423
NYSIIS	.339
DMetaphone	.275
FuzSoundex	.327
Leven dist	.658
Dam-L dist	.659
Bag dist	.597
SWater dist	.889
LCS-2	.915
Skip grams	.844
1-grams	.839
2-grams	.885
3-grams	.783
Pos 1-grams	.890
Pos 2-grams	.880
Pos 3-grams	.768
LCS-3	.909
Compr BZ2	.458
Compr ZLib	.532
Jaro	.853
SAPS dist	.656
SortWink	.803
PermWink	.888
Editex	.631
Winkler	.891
SAPS dist	.656

is measured. In order to choose best suited algorithm for this purpose, we must define how the performance

III. Performance measure

Many new time warping techniques have been developed to improve its computation efficiency. Examples include Index-Based [6], Filter-Based [7], Wavelet [8], Dynamic Time Warping [9], Accounting Causal Relationships [10] and Dynamic Indexing Technique [11] approaches. It is necessary to evaluate algorithms according to the quality of the results of that search using information retrieval measures. In general, Recall and Precision are often used as retrieval effectiveness criteria. According to [12], high recall means retrieving as many relevant items as possible, while high precision means retrieving as few irrelevant items as possible. More specifically, recall is the proportion of relevant matches actually retrieved, and precision is the proportion of retrieved matches which are relevant. A match is relevant if it is judged based on the user interest.

Moreover, having 100 percent precision and 100 percent recall is essential, but it is a challenge. In order to achieve better performance it is necessary to get as maximum precision and recall as possible.

IV. Rule-Based Algorithm

Rule-based algorithms attempt to represent knowledge of common spelling error patterns in the form of various rules for how to transform a misspelled word into a valid one. According to [13], correction candidates are generated by applying all possible rules on the misspelled word and retaining the valid dictionary entries which produces this result and it can be ranked. Frequently, a numerical score is assigned to each candidate, based on the probability of having particular error corrected by the corresponding rule, which means a closer match.

V. Soundex

Soundex, presented in [14], was invented by odell and Russell in 1918 and used by the U.S. Census to match American English names. Soundex translates a name into a four-character code based on the sound of each letter. The first letter of the name is kept constant, while the rest of the letters are coded into digits.

VI. Phonex

Phonex [15] is a variation of Soundex that tries to improve the encoding quality by pre-processing names according to their English pronunciation before the encoding. All trailing s‘ are removed and various rules are applied to the leading part of a name (for example kn‘ is replaced with n‘, and wr‘ with r‘). As in the Soundex algorithm, the leading letter of the transformed name string is kept and the remainder is encoded with numbers (again removing zeros and duplicate numbers). The final Phonex code consists of one letter followed by three numbers.

VII. N-grams

Use an inverted index of n-grams to avoid going through the entire list during search. First, all names in the list are given a unique number,. then for every possible n-gram (with an alphabet of L letters, there are Ln possible n-grams), a list of all the numbers of names containing that n-gram is constructed. In order to find close matches to a specific name, the union of all lists with names having an n-gram in common with that name is taken. Answers can be stored in a heap, sorted after n-gram distance, and the answers with too large distance can be skipped to save sorting time [16].

VIII. Longest common sub-string (LCS)

This algorithm [24] repeatedly finds and removes the longest common sub-string in the two strings compared, up to minimum lengths (normally set to 2 or 3). This algorithmic suitable for compound names that have words (like given- and surname) swapped

IX. Prior studies

Since 1918 several Researches proposes different way to develop English Soundex algorithm, such as Phonix, N-gram, Edit-distance algorithms. At present there are a number of name-matching algorithms employing different degrees of complexity to overcome name variations. Algorithms that use the phonetic features of names have been developed for English, Soundex, presented in [17], was invented by Odell and Russell in 1918 and used by the U.S. Census to match American English names. The Soundex algorithm is designed primarily for English names and is a phonetically based name matching method. Soundex method is more accurate than just relying on character similarities between the names, the Soundex algorithm is not ideal, when comparing first names, different codes could be given for abbreviated forms of a name for example ‗Tom‘ , Thos‘, and Thomas‘ would be classed as different names. Algorithms such as Soundex, Phonix, and Metaphone are all designed for English names. Non-English Phonetic Algorithms are dealing with other languages. Soundex method for French language developed by [18] based on the Russell Soundex method but is adapted for the French language and classifies each name as a three-letter code. Like the Russell Soundx Coding Technique, the names Mireille, Marielle and Merilda which are all given the code MRL. Recent work on improving Soundex focuses primarily on improving performance by manipulating names prior to encoding, or altering Soundex codes after encoding. Examples include Celko‘s [19], Code-Shifting [20], Hodge‘s Phonetex [21] and Editex [22]. Holmes showed that for English, n-gram techniques are less effective than Soundex-based techniques. The explanation provided is that since n-grams are unaware of the phonetic information Soundex uses, they are not able to recognize the phonetic equivalence of various characters. Even so, Hodge notes that n-gram techniques are better equipped to handle insertion and deletion errors. An alternative approach is the use of string distance measures and n-grams. N-gram techniques are language-independent, differing significantly from Soundex in that they do not rely on phonetic similarity. Similarity, as identified by n-grams, is based purely on spelling rather than phonetic information. The two most commonly used values of n are bigrams and trigrams. Edit distances are used to determine the similarity of words after phonetic encoding has been completed. It is significant to note that, unlike other ranking measures, edit distances are not calculated in linear time, given two names, of length p and q, their edit distance would be computed in (pq Θ) and must be computed at run-time, while other techniques such as Soundex and n-grams allow retrieval systems to store encoded names and simply use them at run-time. Personal name matching is very challenging, and more research into the characteristics of both name data and matching techniques has to be conducted in order to better understand why certain techniques perform better than others, and which techniques are most suitable for what type of data. More detailed analysis into the types and distributions of errors is needed to better understand how certain types of errors influence the performance of matching techniques.

X. Methodology

We developed new name matching algorithm, LIPNS, and evaluated it against n-grams and SOUNDEX techniques. The algorithms experimented are as briefly outlined below:

- Prior Work
  - Soundex
  - Phonex
  - N-gram
  - Longest common sub-string(LCS)

- Our Algorithm
  1) LIPNS

1) *Soundex*

Soundex is perhaps the best known and most cited of the similarity key algorithms. Soundex translates a name into a four-character code based on the sound of each letter. The first letter of the name is kept constant, while the rest of the letters are coded into digits according to Table X.1.

Table X.1 soundex phonetic codes

Letters	Code
a ,e, h, i, o, u, w, y	0
b, f, p, v	1
c, g ,j, k ,q, s, x, z	2
d, t	3
L	4
m, n	5
R	6

Letters with the same Soundex digit as their preceding letter are ignored. After coding the entire name, all zeros are eliminated. Finally, the code is truncated or padded with zeros to one initial letter and three digits. As an example Appel → A1104 →A104→A14→A140, Tufaha →T01000→T010→T1→T100.

The encoding algorithm is very fast in practice after the calculation of the code it can be used to quickly lookup possible matches in the name list indexed by Soundex codes. The Soundex algorithm is rather crude and can sometimes go very wrong. Two names with different initial letters will never have the same Soundex code, even though they have the same pronunciation (e.g. Kamel→K540 and camel →C540). The algorithm is designed for English, but even with common English names, it fails easily.

2) *Phonex*

The Phonex Algorithm was first published by Lawrence which is also a phonetic based name matching algorithm. Metaphone algorithm converts a word to any of the combination of the 16 consonant letters. The Conversion rule of Phonex algorithm is like Soundex ignores vowels after the first letter and duplicate letters are not added to the code. It‟s more accurate compared to soundex in certain cases (ex: Bonner and Baymore gives the metaphone codes of BNR and BMR respectively while the Soundex gives the same code which is B560 [25].

3) *N-gram*

There are several different n-gram similarity measures. A simple measure given by [23] is the count of the total number of n-grams two words have in common, gram-count = |N1 \ N2|, where N1 and N2 are the sets of n-grams of the two words. Another measure used by[19], is n-gram distance function, gram-dist = |N1| + |N2| − 2 |N1 \ N2|,|N1| and |N2| denote the number of n-grams in the two words and can be calculated from the length of the words. “Salad” has 4,bigrams, and “Salata” has 5 bigrams. They share three bigrams. The n-gram distance between them is thus 4 + 5 − 2 * 3 = 3, similar example to Arabic product name
“سَلطة” has 3 bigrams and “سَ ا.طا” has 5 bigrams The n-gram distance between them is thus 3 + 5 − 2* 0 = 8 .The similarity measures presented above do not take into account the ordering of letters within words. The two most commonly used values of n are 2 and 3 (bigrams and trigrams).

4) *Longest common sub-string(LCS)*

This algorithm is based on a subroutine computing implicitly the longest common subsequence (LCS) between the text and every substrings of the pattern. This subroutine can be used to compute the length of the LCS between a compressed text and an uncompressed pattern in time O(mn1..5); the same problem with a compressed pattern is known to be NP-hard[26]. For example, the two name strings „gail west‟ and „vest abigail‟ have a longest common sub-string „gail‟. After it is removed, the two new strings are „west‟ and „vest abi‟. In the second iteration the sub-string „est‟is removed, leaving „w‟ and „v abi‟. The total length of the common sub-strings is now 7. If the minimum common length would be set to 1, then the common white space character would be counted towards the total common sub-strings length as well. A similarity measure can be calculated by dividing the total length of the common sub-strings by the minimum, maximum or average lengths of the two original strings similar to Smith-Waterman above). As shown with the example, this algorithms suitable for compound names that have words (like given- and surname) swapped. The time complexity of the algorithm, which is based on a dynamic programming approach [11], is O(|s1|×|s2|) using O(min(|s1|, |s2|)) space [4].

5) *language-independent product name search (LIPNS)*

The LIPNS technique was developed to satisfy the following requirements:

- Product name or any name with small differences should be recognized as being similar.
- If one product name is just a random anagram of the characters contained in the other, then it should (usually) be recognized as dissimilar.

- Language independence- the LIPNS technique should work not only in English, but also in many different languages.

The similarity between two product names is calculated in four steps:

1-Separate product names into letters .

2- Create a matrix by assigning first product name as a row of letters, and second product name as a column of letters.

3- Computing the similarity between letters by assigning one for similar letters and zero for dissimilar letters.

4- Computing the similarity between two product names by using the following formula: Ss(R,C) = 1 - (SumD / L)

- Ss is similarity score
- R is the letters set for the first product name (Row)
- C is the letters set for the second product name (Column)
- SumD is the summation of the ones lies on diagonal matrix.
- L is the length of product name in the DB

Assume that all relation scores are in the {0, 1} range, which means that if the score gets a minimum value (equal to 0) then the two product names are absolutely similar. To obtain effective results, the user has to just increase/decrease the Score value Estimator, which was estimated at (0.25), this score value was obtained through repeated trials and strenuous efforts based on the user‟s terminal benefit and satisfaction as main consideration. For example, the LIPNS for "Salata" and "Salad" are shown below according to

Table X.2.

Table X.2 Similarity Matrix for „Tomato‟ and „Tamato‟

	T	o	m	a	t	o
T	1	0	0	0	1	0
a	0	0	0	1	0	0
m	0	0	1	0	0	0
a	0	0	0	1	0	0
t	1	0	0	0	1	
o	0	1	0	0	0	1

Ss = 1 – (5 / 6)

Ss= 0.16

Since the Ss result is less than 0.25 , „Tomato‟ and „Tamato‟ are considered to be similar. In order to improve the performance we modified LIPNS steps as follow:

M.1- Compare two names before step 1, if two names are similar then stop and exit (matching).

M.2- If M.1 not matching then go to step one and two (LIPNS).

M.3- Delete not matching letter from both names and stop matching

M.4 -Go back to M.1 and continues.

If number of letter elimination is more than two, both names are not similar and stop matching. For example, the MLPINS similarity for "Tomato "and "Tamato "are performed as follow.

If "Tomato " = "Tamato then They are Similar , stop MLIPNS.Else Build matrix

	T	o	m	a	t	o
T						
a		~~D(stop)~~				
o						

If "Tmato " = "Tmato then Similar and stop MLIPNS.Else Continue matrix building and starting from next letter where you stop before (in this case m)

Etc...

XI. Result

The results show that R-precision is superior to average precision for the Arabic product name search task since the number of relevant result is too small for interpolation. However, weak ordering in the result sets demands that R-precision be calculated over a number of random permutations of the results. The LIPNS, Soundex and N-gram were tested over queries and the effectiveness of the result is shown in Table XI .1.

Table XI .1 Comparison of various name matching algorithms with LIPNS

Technique	Average precision	R-Precision
Soundex	0.39089	0.2451
Bigrams	0.3339	0.1823
Trigram	0.13772	0.01151
LIPNS(Ss=0.25)	0.7579	0.5970
LIPNS(Ss=0.3)	0.3925	0.2591

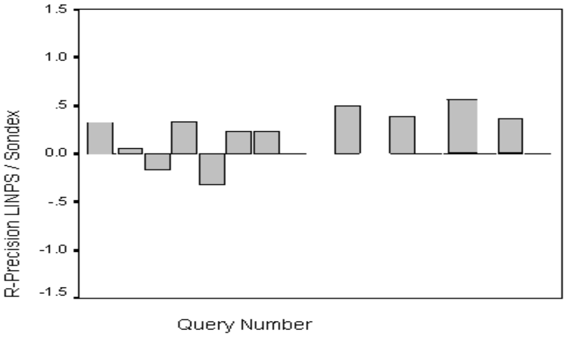
Table XI.2 Comparison of various name matching algorithms with MLIPNS

Technique	Average precision	R-Precision
Phonex	0.69	0.49
LCS	0.73	0.52
MLIPNS	0.81	0.95

To confirm the significance of these results ,we collected the results from all five of the scenarios described above and tested their composite significance, by using the R-precision measures for several queries to compare the retrieval history of two algorithms as follows. RPLipns/Soundex(i)=RPLIPNS (i)– RPSoundex(i). A value of RPLipns/Soundex(i) equal to 0 indicates that both algorithms have equivalent performance (in term of R-precision) for the i-th query. A positive value of

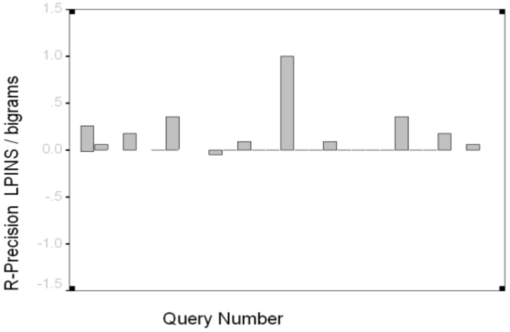
RPLipns/Soundex(i) indicates a better performance by algorithem LINPS (for the i-th query) while a negative value indicates a better retrieval performance by algorithm Soundex. For instance, the difference RPLipns/Soundex(i)= 0.5970 - 0.2451 is 0.3519, this indicates a better performance by algorithem LIPNS. Figure XI.1 illustrates the RPLipns/Soundex(i) values (labeled R-Precision LIPNS / Soundex) for two hypothetical retrieval algorithms over eleven sample queries. The algorithm LIPNS is superior for nine queries while the algorithm Soundex performs better for the two other queries.

Figure XI.1 Precision histogram for eleven hypothetical queries



The difference RPLipns/bigrams(i)= 0.5970 - 0.1823 is 0,4147, this indicates a better performance by algorithm LIPNS. Figure XI.2 illustrates the RPLipns/bigrams (i) values (labeled R-Precision LIPNS / bigrams) for two hypothetical retrieval algorithms over eleven sample queries. The algorithm LIPNS is superior for ten queries while the algorithm bigrams performs better for the one other query.

Figure XI.2 Precision histogram for eleven hypothetical queries



In other hand, trigram doesn‘t perform better for any queries, compared to LIPNS algorithm. From these result, we may draw the fact that LIPNS provides a statistical significant improvement over both Soundx and bigrams in the general situation, also obvious is the fact that our LIPNS and MPLINS techniques provide an improvement in performance that is significant even at the 99% level. The result obtained through this new method is purely independent of languages and different type of characters in Arabic, Latin, Indian, Russian etc. LIPNS and MLIPNS is an innovative method of its kind which is totally independent of all the world languages and a globally long awaited concept. The outcome of this new method clearly shows that this method should be superior to any other earlier methods available.

XII. Conclusions

The new method developed is purely independent of languages and different type of characters in Arabic, Latin, Indian, Russian etc. has been experimented. The SOUNDX test works only for English characters and names, whereas the proposed method has proved that it is not restricted to English language only and outperform over the other methods. It also compares LPINS method with n-gram method and proved that newly proposed method is superior to n-gram method as well. The specialty of LPINS method is that it works with all languages as well and it is an efficient method to implement for any related application. As per the formula derived here in this study, the researcher or user can control and modify the score value which will affect the R-precision and Average precision value. This is something new and does not exist in other methods according to our knowledge. The new formula found in this study can be controlled and easily modified to adjust the score values as it gives a user friendly environment. For example, to obtain effective results, the user has to increase/decrease the Score value Estimator, which was estimated here is at 0.25. This score value was obtained though repeated trials and strenuous efforts based on the user‘s terminal benefit and satisfaction. The score value has been selected to be main criteria to measure the similarity between the product names. Even though, this new method performs better, this can be applied with different applications such as name verification in customer data bases. Execution times are very important for a matching process to take place and show the results especially when the matching is attached to a sequence business process, in order to achieve these objectives we develop MLIPNS technique. Hence an appropriate decision on the algorithm to be used should be decided based on the algorithm accuracy, quality of data and Execution time of the algorithm. The appropriate algorithm can be decided by optimizing these three factors based on the Business requirements.

XIII. References

1) Ted Friedman, Andreas Bitterer, ” Similarity Buys Evoke to Gain Technology Market Presence” , © 2005 Gartner, 20 July 2005 ID Number: G00129921.

2) F. Patman and P. Thompson. Names: A new frontier in text mining. In ISI-2003, Springer LNCS 2665, pages 27–38

3) A. J. Lait1 and B. Randell. ”An Assessment of Name Matching Algorithms” Department of Computing Science University of Newcastle upon Tyne,2005.

4) Peter Christen. A Comparison of Personal Name Matching: Techniques and Practical Issues”, Computer Sciences Laboratory Research School of Information Sciences and Engineering, September 2006.

5) Chakkrit Snae, Acomparison and Analysis of Name Matching Algorithms", International Journal of Applied Science, Engineering and Technology Volume 4 Number 1 2007 ISSN 1307-4318.
6) Kim S W, Park S, and Chu , Efficient processing of similarity search under time warping in sequence databases", Information Systems 29: 405–20, 2004.
7) Kwak T Y and Lee Y J A filtering method for searching similar multi-dimensional sequences under the time-warping distance", Information Systems 28: 791–813, 2003.
8) Chan F K P, Fu A W C, and Yu C ," Wavelets for efficient similarity search of time series With and without time warping", IEEE Transactions on Knowledge and Data Engineering 15: 686–705, 2003.
9) May Yuan, John McIntosh, Assessing Similarity of Geographic Processes and Events", © Blackwell Publishing Ltd., 2005.
10) Ella Mae Matsumura, Sandra C. Vera-Muñoz, ApplyingAccounting Causal Relationships: Experimental Evidence on the Decision Performance Effects of Problem Similarity and Comparison", Wisconsin Alumni Research, 2005.
11) G. Qian, Q. Zhu, Q. Xue and S. Pramanik , A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces", ACM, 2006.
12) Ricardo Baeza-Yates & Berthier Ribeiro-Neto Modern Information Retrieval", Publishers, New York, Addison-Wesley, 1999.
13) P. A. V. Hall and G. R. Dowling. Approximate string matching", ACM Com-putting Surveys, 12(4):381–402, 1980.
14) Camps, R., & Daude, J. Improving the efficacy of approximate personal name matching", 8th International Conference on Applications of Natural Language to Information Systems, 2003.
15) Gadd, T.N. PHONIX: The Algorithm", Program – Electronic Library and Information Systems, 1990.
16) Pfeifer, U., & Poersch, T., & Fuhr, N. Searching Proper Names in Databases". Proceedings of the Conference on Hyper-text – Information Retrieval – Multimedia, Germany, pages 259-275, 1995.
17) Binstock, A., & Rex, Practical Algorithms for Programmer". Reading, MA: Addison-Wesley, 1995.
18) Gerard Bouchard and Christian Pouyez, Name Variations And Computerized Record Linkage", Historical Methods, Vol. 13, No. 2, pp119-125, Spring 1980.
19) Celko, J. Joe Celko's SQL For Smarties: Advanced SQL Programming", 2nd Ed., Burlington, MA: Morgan Kauffman, 1995.
20) Holmes, D., & McCabe, M. Improving Precision and Recall for Soundex Retrieval", Proceedings of the 2002 IEEE International Conference on Information Technology - Coding and Computing, pages 22-28, 2002.
21) Hodge, V., & Austin, J. Technical Report YCS 338", Department of Computer Science, University of York, 2001.
22) Zobel, J., & Dart, P. Phonetic String Matching: Lessons from Information Retrieval", Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 166-172, 1996.
23) Pfeifer, U., & Poersch, T., & Fuhr, N." Retrieval Effectiveness of Name Search Methods" Information Processing and Management, 32(6), pages 667-679, 1996.
24) C. Friedman and R. Sideli ,"Tolerating spelling errors during patient validation", Computers and Biomedical Research,25:486–509, 1992.
25) Rupesh Shyamala ,"Name Address Matching Matching: A F Fuzzy perspective using various Name Matching Algorithms", 2006.
26) Alexander Tiskin, Faster subsequencet recognition in compressed strings", Department of Computer Science, The University of Warwick,Coventry CV4 7AL, United Kingdom,2006.

An Empirical Comparison of: HTML, PHP, ColdFusion, PERL, ASP.NET, JavaScript, VBScript, PYTHON and JSP

Amadin I.F.[1] , Nwelih E.[2]

***Abstract*-With the advent of the World Wide Web, several Web development languages have emerged and selecting a suitable one is never an easy task. Over the years, several attempts have been made to evaluate Web development tools vis-à-vis software measurement. This paper presents an experimental evaluation of nine web development languages. A shopping cart application was implemented in each of the Web development languages and the following factors were used in our evaluation: Platform, Performance, Functionality, Ease of use, Reliability, Program length, Portability, Database supports, Speed of execution, Maintainability, Object oriented programming, and Development cost.**

Keywords-Web Development languages, World Wide Web, Information Technology (IT) and Web browser

I. Introduction

The World Wide Web (WWW) has rapidly become the standard for displaying information on the Internet. Over the last ten years, the nature of the Internet and the World Wide Web has changed drastically from what they were then. Exciting new developments occur almost daily, as the pace of innovation is unprecedented by any other technology (Berner-Lee, 2002).This tremendous advancement is mainly due to improvements in the design of Web sites with the aid of sophisticated Web development technologies and languages. The advent of these languages have changed the content of the Internet from its usual Web pages full of only static text and very few images to web sites that not only animate text and images but offer a wide range of services including database and multimedia features. Terms like portals, e- (commerce, payment, education, banking, learning etc) have become everyday terms due to the improvement in Web development technologies and their familiarity with the general public. Several methods and experiments have been carried out to determine the suitability and effectiveness of Web development tools which have led to the emergence of three evaluation methods namely: Empirical, Vendor and Usability. The metrics for comparing Web development tools coding and design representations is reported in Ovum (2000). Work on assessing an aspect of a visual Programming Language (writing matrix multiplication problems) is reported in Pandey and Burnett (1993) while Apte and Kimura (1993) examined the relative merits of two input devices for editing graphic diagrams. By far, the most extensive and ambitious research conducted using empirical comparison was done by Prechelt (2000). Though the work was done by making use of a phone code program written in each of the seven programming Language or Web technologies, the author concluded that the work was not conclusive enough in terms of serving as a comparison guide. The growth of the Web is phenomenal, but the number of Web users is now measured in the tens of millions while the number of Web sites is now measured in the millions. Regardless of the actual numbers, it is clear that Corporate Companies, Academic Institutions, Government have spent a lot of time, attention, and money on the Web, Research Organisations etc wanting to get involved in Cyberspace. Very few of them have much of a feel for their payback on this investment. Much of that has been due to the incredible hype and fast growth surrounding this technology, combined with the low cost of experimentation with the latest and emerging sophisticated Web development tools or technologies available in the software market that suit their need.

II. Background

A programming language is simply referred to as a system of communication with its own set of conventions and special words used to interact with the computer system. A programming language enables a programmer to dictate what, how and when a computer system will perform a task. In this modern age of information technology (IT), where the computer and the internet has now become a key player in every area of our lives, the need for a comparative study on the different languages used to interact with the computer and internet has become necessary. According to Janstal (2000), Web development tools are failing to address users' needs despite the promises made by vendors. However, the market has no clear leader, and there are inadequate products for medium or large-scale development projects, (Ovum, 1997). According to Mahar (1997), the main deficiencies in web development tools are that they cannot support teams of developers working together. In this fast evolving area, according to Ward-Dutton (2002) in his works, he analysed a range of tools for building integrated Web applications. Developers are continuously in search of tools or technologies available in the software market that suit their need, as thousands of Web development tools or technologies exist. Web development technologies are equally subject to the Laws of Evolution (survival of the

About[1]-Department of Computer Science Faculty of Physical Sciences University of Benin P.M.B 1154, Benin City Nigeriafrankamadin@yahoo.com

About[2]- Department of Computer ScienceFaculty of Physical Sciences University of Benin P.M.B 1154, Benin City Nigeriaemmanuelnwelih@yhoo.com

fittest) and there are some criteria used in measuring their acceptability and usability. To measure the usability of Web development tools or technologies which are particularly useful for the development of dynamic Web pages and animated movies, Green and Petre (1996) introduced the popular Cognitive Dimensions framework which is a broad-brush evaluation technique for interactive devices and for non-interactive notations. It sets out a small vocabulary of terms designed to capture the cognitively-relevant aspects of structure, and shows how they can be traded off against each other. The development and application of various metrics for comparing visual and textual representations is reported in Nickerson (1994). Some common web development tools are discussed below:

HTML:Hyper Text Markup Language (HTML) which was founded in 1980 by Tim Berners-lee, is the predominant markup language for web pages. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists etc as well as for links, quotes, and other items. It allows images and objects to be embedded and can be used to createinteractive forms. It is written in the form of HTML elements consisting of "tags" surrounded by angle brackets within the web page content. (Tim, 2001)

PHP:Hypertext Preprocessor (PHP), was conceived in 1994 by Rasmus Lerdorf. He wrote it as a way to track visitors to his online CV. The first version was released in early 1995; Rasmus had found that by making the project open-source, people would fix his bugs. The first version was very straightforward and had a simple parser which recognized a few special macros and provided some of the utilities which were in common usage on web-pages back then. In 1995 it was renamed PHP/FI version 2. The "FI" in this version stood for the Form Interpreter which he added to PHP to cope with the growing needs of web-pages and mSQL (Microsoft SQL) support was also added at this time. PHP/FI underwent massive growth, and other people started to contribute code to it regularly. In 1997, Zeev Suraski and Andi Gutmans rewrote the main parser, and PHP shifted from being Rasmus' own to a more group orientated project. This formed the basis for PHP3, now named PHP: Hypertext Preprocessor with version such as PHP4 and PHP5 engine. The latest version comes with most of those features which were not in the earlier versions of PHP. It is easier to integrate into existing HTML pages, faster and more efficient for complex programming tasks and trying out new ideas. PHP is generally referred to as more stable and less resource intensive as well. (Valade, 2004)

COLDFUSION:ColdFusion (CF) is an application server and software language used for Internet application development such as for dynamically-generated web sites. In this regard, ColdFusion is a similar product to Microsoft Active Server Pages, Java Server Pages or PHP. ColdFusion was the first amongst these technologies to provide the developer the capability of creating dynamic websites that were attached to a backend database i.e. Cold Fusion has better database abstraction. The primary distinguishing feature of ColdFusion is its associated scripting language, ColdFusion Markup Language (CFML), which compares to Active Server Pages, JSP, or PHP resembles HTML in syntax. "ColdFusion" is often used synonymously with "CFML", but there are additional CFML application servers besides ColdFusion, as ColdFusion supports

programming languages other than CFML, such as server-side Actionscript and embedded scripts that can be written in a JavaScript-like language known as CFScript. ColdFusion was originally developed by brothers JJ and Jeremy Allaire in July 1995. In 2001 Allaire was acquired by Macromedia, which in turn was acquired by Adobe Systems in 2005. ColdFusion is most often used for data-driven web sites or intranets, but can also be used to generate remote services such as SOAP web services or Flash remoting. It is especially well-suited as the server-side technology to the client-side Flex. ColdFusion can also handle asynchronous events such as SMS and instant messaging via its gateway interface; it also has a good error handling capability, date parsing features and more. Cold Fusion is available on Win32, Solaris, Linux and HP/UX operating systems respectively. Published by Allaire in 2004, ColdFusion 4.0 is an enterprise level Web application development suite. This means it can be used not just to develop simple Web pages but also to develop databases, and more dynamic Web sites. The product is now in its fourth version and has been a consistent market leader. ColdFusion uses its own anatomized scripting tags, which when embedded in a Web page are read by the Web server. The Web server then produces dynamic output for the end user. All ColdFusion scripting is browser independent, making its content available to a wide audience. Only in the exception that CFML is combined with DHTML will a high end browser be needed. Instructions are passed to ColdFusion using templates. A template looks like any HTML file, and the only difference being the CFML tags. (http//www.adobe.com/products/coldfusion. Retrieved, 2010)

PERL: The first version was introduced in the year 1987 by Larry Wall. The author’s purpose for the creation was as a result of the disappointing result of languages such as sed, C, awk and the Bourne Shell offered him. He looked for a language that will combine all of their best features, while having a few disadvantages of its own. Since then, Perl has seen several versions of each additional function. Perl version 5, which was released in 1994, was a complete rewrite of the Perl interpreter, and introduced such things as hard references, modules, objects and lexical scoping. Several minor versions of Perl appeared since then, and the most up-to-date stable version (as ofOctober 2005) is 5.8.x. Perl became especially popular as a language for writing server-side scripts for web-servers. But that's not the only use of perl, as it is commonly used for system administration tasks, managing database data, as well as writing GUI applications. One problem with the pearl language is its flexibility / complexity that makes it easier to write code that another author / coder has a hard time reading. (Rice's Theorem, 2008; Wikipedia, 2010)

ASP.NET:The Active Server Page (ASP.NET) was co-developed by Mark Anders, a manager on the IIS (Internet Information Server) team, and Scott Guthrie, who had joined Microsoft in 1997 after graduating from Duke University. The initial design was developed over the course of two months by Anders and Guthrie, and Guthrie coded the initial prototypes during the Christmas holidays in 1997. ASP.NET is a web application framework developed and marketed by Microsoft to allow programmers to build dynamic web sites, web applications and web services. It was first released in January 2002 with version 1.0 of the .NET Framework, and is the successor to Microsoft's Active Server Pages (ASP) technology. ASP.NET is built on the Common Language Runtime (CLR), allowing programmers to write ASP.NET code using any supported .NET language such as VB.NET, C#, VC++.NET, etc. ASP.NET pages, known officially as "web forms", are the main building block for application development. Web forms are contained in files with an ".aspx" extension; in programming jargon, these files typically contain static (X)HTML markup, as well as markup defining server-side Web Controls and User Controls where the developers place all the required static and dynamic content for the web page. Additionally, dynamic code which runs on the server can be embedded within webpages within a block <% -- dynamic code -- %> which is similar to other web development technologies such as PHP, JSP, etc. The biggest drawback of ASP is that it's a proprietary system that is natively used only on Microsoft Internet Information Server (IIS). This limits it's availability to Win32 based servers. (http://www.asp.net/ Retrieved, 2010)

JavaScript: JavaScript is an object-orientedscripting used to enable programmatic access to objects within both the client application and other applications. It is primarily used in the formof client-side JavaScript, implemented as an integrated component of the web browser, allowing the development of enhanced user interfaces and dynamic websites. JavaScript is a dialect of the ECMAScript standard and is characterized as a dynamic, weakly typed, prototype-based language with first-class functions. JavaScript was influenced by many languages and was designed to look like Java, but to be easier for non-programmers to work with. JavaScript was originally developed by Brendan Eich of Netscape under the name *Mocha*, which was later renamed to *LiveScript*, and finally to JavaScript. The change of name from LiveScript to JavaScript roughly coincided with Netscape adding support for Java technology in its Netscape Navigatorweb browser. JavaScript was first introduced and deployed in the Netscape browser version 2.0B3 in December 1995. The naming has caused confusion, giving the impression that the language is a spin-off of Java, and it has been characterized by many as a marketing ploy by Netscape to give JavaScript the cachet of what was then the hot new web-programming language. JavaScript, despite the name, is essentially unrelated to the Java programming language even though the two do have superficial similarities. Both languages use syntaxes influenced by that of Csyntax, and JavaScript copies many Java names and naming conventions. The language's name is the result of a co-marketing deal between Netscape and Sun, in exchange for Netscape bundling Sun's Java runtime with their then-dominant browser. The key design principles within JavaScript are inherited from the Self and Scheme programming languages.

VBScript: Visual Basic Scripting (VBScript) is an Active Scripting language, developed by Microsoft, which uses the Component Object Model to access elements of the environment within which it is running (e.g. FileSystemObject or FSO used to create, read, update and deletefiles). The language's syntax reflects its origins as a limited variation of Microsoft's Visual Basic programming language. VBScript has been installed by default in every desktop release of Microsoft Windows since Windows 98; as part of Windows Server since Windows NT 4.0 Option Pack; and optionally with Windows CE (depending on the device it is installed on). VBScript script must be executed within a host environment, of which there are several provided with Microsoft Windows, including: Windows Script Host (WSH), Internet (IE), Internet Information (IIS). Additionally, The VBScript hosting environment is embeddable in other programs, through technologies such as the Microsoft Script control. VBScript began as part of the Microsoft Windows Script Technologies, which were launched in 1996, initially targeted at web developers. During a period of just over two years, theVBScript and JScript languages advanced from version 1.0 to 2.0, and over that time it gained support from Windows system administrators seeking an automation tool more powerful than the batch language first developed in the late 1970s. In version 5.0, the functionality of VBScript was increased with new features such as: regular expressions; classes; the *With* statement; the *Eval*, *Execute*, and *ExecuteGlobal* functions to evaluate and execute script commands built during the execution of another script; a function-pointer system via GetRef, and Distributed COM (DCOM) support. In version 5.5, *SubMatches* were added to the *regular expression* class in VBScript, to finally allow VBScript script authors to capture the text within the expression's groups. That capability before was only possible through JScript. With the advent of the .NET framework, the scripting team took the decision to implement future support for VBScript within ASP.NET for web development, and therefore no new versions of the VBScript engine would be developed and it moved over to being supported by Microsoft's *Sustaining Engineering Team*, who are responsible for bug fixes and security enhancements. For Windows system administrators, Microsoft suggests that they migrate to Windows PowerShell. However the scripting engine will continue to be shipped with future releases of Microsoft Windows and IIS.

PYTHON:Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. The programming language was conceived in the late 1980s by Guido Van Rossum at CWI in the Netherlands as a successor to the ABC programming language (itself inspired by the SETL) . Ever since, various versions of the language has emerged with new features ported with the various versions of the language. Python is a multi-paradigm programming language. This means that, rather than forcing programmers to adopt a particular style of programming, it permits several programming style: object oriented and structured programming are fully supported. The python language has a very good memory management capability. http://www.python.org/about/ Retrieved, 2009)

JSP: Java Server Pages (JSP) is a Java technology designed by Sun Microsystems that allows software developers to create dynamically-generated web sites, with HTML, XML, or other document types, in response to a Web client request. The technology allows Java code and certain pre-definedactions to be embedded into static content. The JSP syntax adds additional XML-like tags, called JSP actions, to be used to invoke built-in functionality. Additionally, the technology allows for the creation of JSP tag libraries that act as extensions to the standard HTML or XML tags. Tag libraries provide a platform independent way of extending the capabilities of a Web server. The JSP engine/compiler (An engine that compiles codes written in the java server pages) is built around the Servlet Engine i.e. JSP is servlet made easy as most of the complicated task in servlet were made easy. JSPs are compiled into Java Servlets by a JSP compiler. Which may generate a servlet in Java code that is then compiled by the Java compiler, or it may generate byte code for the servlet directly. JSPs can also be interpreted on-the-fly, reducing the time taken to reload changes. Even though JSP is platform • independent, it requires the coder/programmer to be familiar to the java language very well or the object oriented programming technique. (Java.sun.com/products/jsp. Retrieved, 2010)

III. Materials and Methods

Very often technology based decisions are made by technical personnel who base their decision on personal use, attendance at vendor sponsored workshops, reading about it in trade publications or having used other products from the same vendor. In this work, the empirical approach was used to evaluate the Web development tools under consideration. coding, compiling / interpreting and running programs in each Web tools.

1) *Ease Of Use*

Is the ease which the language is used in developing an application and the availability of structures that reduce programming complexity. For instance, some Web development tools with support for GUI are more users friendly and aid usability than Web tools without those features.

Ease of use: would be determined by taking note of how easier it is to write or design programs using any of the selected Web tools.

2) Speed Of Execution

The time it takes to compile and execute. The amount of time is measured by use of a stopwatch or by building in some program segment (s) to keep track of the execution and compilation time.

Speed of execution: this is obtained by a program module that records start and stop time of execution.

3) *Platform*

This described the ability of software to run on a variety of different operating systems or the same operating systems. Different operating systems provide different platform challenges for web development tools. Issues with respect to 32 bit and 64 bit operating system are prominent.

Platform: the ability of the program to run on the same or different operating systems and hardware.

4) *Functionality*

This described the ability of software to function properly or meet users' needs in order to achieve their desired goals.

Functionality: entails that the Web tools respond to users' need in order to achieve their desire goals without any delay in delivery the Web content.

5) *Performance*

This described how will the Web development technologies can be used to achieve quickly and efficient delivery of applications.

*Performance:*can be determined by considering how quickly and efficiently the selected tools are used to implement a Web site.

1. *Program Length*

Program length is the same as the number of lines of codes present in the program which contains anything that contributes to the semantics of the program in each of the program files e.g., a statement, a declaration or at least a delimiter such as a closing brace or tags.

Length: manually counting the number of executable lines of codes for the implementation.

2. *Program Reliability*

It is concerned with how well programs behave. Web development tool reliability entails that the tool must consider unforeseen errors like syntax errors and other forms of language violations and responds appropriately by informing the user(s) of any violation instead of terminating executions inadvertently.

Reliability: measured by Web page output presentation in response of the Web browser / Web server.

3. *Development Cost*

The cost of developing a program in the target Web development tool is determined by the total man hours used in bringing out the final version of a Web application program, the cost in terms of the systems resources used and procurement of software and any other device or resource.

Development cost: would involve a collection of all expenses incurred in getting each of the selected Web tools, the machine time utilized in terms ofweb browser

*Functionality***:** entails that the Web tools respond to users' need in order to achieve their desire goals without any delay in delivery the Web content.

5) *Performance*

This described how will the Web development technologies can be used to achieve quickly and efficient delivery of applications.
*Performance:*can be determined by considering how quickly and efficiently the selected tools are used to implement a Web site.

6) *Maintainability*

This described how easy the Web development tools could adapt to changes when the need arises.
Maintainability: can be determining by considering how debugging is carried out when there is an error in the program and modifying tosuit required upgrade.

7) *Object Oriented Programming Design Facilities*

This described the ability for the Web development tools to support OOP, which enhances reuse of object and quicker way to develop application.
Object-oriented programming facilities: determined by considering the various Web tools if they have the ability to used objects for programming reusability.

8) *Database Supports*

This described how the Web development technologies support a wide variety of back-end databases for effective records keeping.
Database supports: tests compatibility by linking the Web tools to a variety of database programs.

9) *Portability*

A term applied to software that is not dependent on the properties of a particular machine, and can therefore be used on any machine. Such software is also described as portable.
Portability: can be determined by implementing these tools on various computer machines in order to know whether it is machine dependent or independent

III. Results and Discussion

Empirical Evaluation of Web development tools
Using the identified 12 criteria to evaluate the algorithm of each tool to ascertain their worthiness based on application developed with them. A simple shopping cart program was implemented with the case tools and the results obtained alongside the apparent conclusions are given as follows:

1) *Program Length*

The number of lines of code for each algorithm
implemented for Web development tool is as shown below:

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Observable number of line of codes	10	8	8	8	8	10	10	10	5

Table 1: Average Program Length of the different Web development technologies
An examination of the data in Table 1 shows an increase in the source codes.

2) *Program Reliability*

Reliability of programs written in any programming languages/Web development technologies is never easy to determine as there are different parameters used by different software practitioners. In our owncase, we considered the ease with which each of the Web development tools implemented their source code, their response when no input was fed in and the response of the interpreter to syntax errors. HTML and PHP were very reliable while the others were reliable.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Execution and interpreter behaviour	Interprets well, error messages display and performs poorly when empty data are encountered	Interprets well, but slowly; terminates execution when empty data are encounter-ed error messages not display	Interprets and execute well, error messages not display	Interprets well and error messages display	Interprets well and error messages display	Interprets well and error messages display	Interprets well and error messages display	Interprets well and error messages display	Interprets well and error messages display
Reliability rating	Very reliable	Very reliable	Reliable	Reliable	Reliable	Reliable	Reliable	Reliable	Reliable

Table 2: Reliability of the selected Web development technologies

3) Development Costs

The combined cost of acquiring an interpreter, setting it up, and that of thesystems resources taken up by the Web page design in the target Web development technologies are summarized below in Table 3.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Average cost of interpreter ($)	Free	Free	80	300	400	350	180	160	Free
Average set up costs ($)	Free	Free	2.40	4.50	6.80	3.20	2.70	2.64	5.40
Systems Requirements	1,064,356	1,024,543	1,004,564	1,324,097	1,423,206	1,300,340	1,320,543	1,375,300	1,400,200
Costs of writing Programs/computer time	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50

Table 3: Development Costs of the different Web development technologies

4) Ease Of Use

Rating the easy to use of the Web development technology with a scale of 1 to2 (where 1denotes easiest to learn and 2 denote easy). Table 4 shows that HTML, PHP, CF, PERL, and JSP are the easiest to learn, while ASP, JS, VBS and PYTON are easy to use.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Easy to use (When scaled)	1	1	1	1	2	2	2	2	1

Table 4: Ease of Use of the selected Web development technologies

5) Speed Of Execution

The speed of execution (measured inseconds) for the simple Web pagecreated is given below in tables

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Interpreting/speed of execution (in seconds)	0.56	0.60	0.60	0.60	0.60	0.58	0.58	0.58	0.60

Table 5: Average Speed of Execution of the different Web development technologies

6) Platform

Testing each tool on two major operating systems; Windows Operating System 98 and Windows XP Operating system determined the platform supports by theWeb development tools examined. From our observation, it was apparent to us, to scale the tools into the platform they support: 1 for dependent; 2 for independent. Table 6 shows that all the programming languages are independent.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Platform (When scaled)	2	2	2	2	2	2	2	2	2

Table 6: Platform Support of the different Web development technologies

7) Funcitonality

Functionality of the Web development tools can measure by the way the tools meets users need based on the fact that they can easily be used to designed sophisticatedWeb page or Web application. Thus, the functionality can also be rated as 1 for excellent, 2 for very good, 3 for good, and 4 for poor.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Functionality (When scaled)	1	2	2	2	2	2	2	2	2

Table7: Functionality of the different Web development technologies

8) Maintainability

Maintainability had to be measured by use of established methods for determining the effect of modifying or debugging of errors from the coded Web development tools. It was observed that HTML, PHP, CF, and PERL were very easy to maintain whileASP, JS, VBS, PYTON, and JSPwere easy to maintain. From the general point of view, we can conclude that the maintainability of Web tools is okay.

Rating the maintainability of the Web development tools with a scale of 1 to 4 (where 1 stands for very easy, 2 for very difficult, 3 for easy, 4 difficult) would give us the following table.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Maintainability (When scaled)	1	1	1	1	3	3	3	3	3

Table 8: Maintainability of the different Web development technologies

9) Object Oriented Programming Facilities

This can be determined by the usage of its features in the various Web developmenttechnologies. From our observation, we can rate the various tools into the classes they belong. That is, 1 for excellent, 2 for very good, 3 for good, and 4 for poor.

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
OOP (When scaled)	1	4	4	4	2	2	3	3	4

Table 9: OOP Facilities of the different Web development technologies

Determining the development costs for the various language implementations was a bit difficult and we had to arrive at a compromise to base the actual costs of the interpreters in US Dollars since that is the most recognized currency used in international and on-line business transactions. The costs of the interpreters were found to vary from one marketer to the other and after comparing prices for 8 (eight) different retailers, we took the average price which was in the range of prices offered and which included the prices for shipping. The set up costs indicated in the offer prices were used while for the cost of writing, coding and running each algorithm in the selected languages, a flat rate of $1.50 per hour of computer time was assumed if one were to carry out the programming task in a commercial center. The assumed price is closest to the N250 charged per hour of computer time in business centers. An examination of table 9 shows that the development cost of the HTML program is the cheapest while the development costs of the Macromedia ColdFusion program is the most expensive.

10) Database Supports

The database supports is determined by the linkage of the Web page to some of the various database software. Form our observation, the following rating was used to determined the levy of support to the database, that is, (1 for excellent, 2 for very good, 3 for good, and 4 for poor

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Database supports (When scaled)	1	4	4	4	2	2	2	2	2

Table 10: Database supports of the different Web development technologies

11) Portability

The portability of the Web development tools based on the simple Web page designwith then was measured based on the execution of the program in different computer machines. From our observation, we can rate the portability into scale as (1for highly portable, 2 for portable, 3 forfairly portable and 4 for not portable).

Technology	HTML	PHP	CF	PERL	ASP	JS	VBS	PYTON	JSP
Portability (When scaled)	1	1	1	1	2	2	1	2	1

Table 11: Portability of the different Web development technologies

IV. Conclusion

From our findings in this research, we conclude that the choice of users to a particular web tool is based on the task at hand which may be designing a simple webpage:- Use HTML, FrontPage editor or PHP; designing for efficient database support:- PHP, CF, ASP, JS, VBS, and PYTON; designing for Object oriented task:- Use PHP, Java Script; designing for Low cost:- PHP, HTML, ASP, VBS, JS; designing for small program length:- JSP since it requires less coding; designing for portability:- PHP, HTML, PYTON, and JS; designing for functionality:- PHP, JSP and XML and designing for Speed of Execution:- PHP. This shows that the use of reliable approaches or software development processes, copious analyst, good designers, and painstaking implementation techniques are a pre-requisite. Thus, very rich design and coded Web site for transactions in the Internet produces attractiveness, and user-friendly Web pages.

V. Recommendation

Web development technologies have become the core of most organization, Corporation, Research Institute, Government presence on the Internet, it is imperative that the use of these tools be accorded the attention it desires. One way of according this attention is for organization, corporate companies, research institutions and Government to acquire these sophisticated tools and trained their IT staff on how to use the tools to develop dynamic and interactive Web sites or custom applications that will ease this function. We strongly recommend that individual or organizations should select their Web development tools based on the level of usability, which strongly encompasses other factors. Universities do not just produce graduates strictly for academic research but also for industrial purposes, therefore, there is the need for Lecturers to be well acquainted with these tools in other to empower their students for job opportunities after graduation

Who will benefit from this report?

IT managers and strategists who need to assess whether they should develop business applications for the Web, and if so, when and how to do it

Software developers who need to identify the right tool for their projects

Business managers who need to understand how the Web can make their business processes simpler and more efficient

Applicationdevelopment tool vendors who need to identify opportunities for partnership or assess competing products

Consultants and systems integrators who need to advise their clients on the variety of different approaches to intranet and WWW development

VI. References

1) Active Server Pages : "An Introduction to Web-based Application Development" available at http://www.abiglime.com/webmaster/articles/asp/122297.htm
2) Bakken S.S., (2000), Introduction to PHP", available at http://www.zend.com /zend/hop/tas,as.php
3) Bakken S.S., (2000), A Brief History of PHP", available at http://www.php.net/ manual/en/intro-history.php
4) Berner-Lee, T., (2002), The World Wide Web – Past Present and Future available at http:// uazuay.edu.ec/bibliotecs/mbaTi/pdf/The%20

World%20Wide %20Web%20Past%20Present%20 and %20Future%20TimBL.pdf

5) Bloor Research Group: Web Based and Client Side Development tools: An Evaluation & Comparison available at http://dpu.se/bloweb-e.html-27k
6) Brajnik, G., (2000), "Automatic Web Usability evaluation; what needs to be done?" 6th conference on Human factors & the Web available at http://www.citeseer.ist.psu.edu/context/148084/0-12k
7) Deitel, H.M., Deitel, P.J., and Goldberg, A.B., (2004), "INTERNET & WORLD WIDE WEB How to program", 3rd Edition, Published by Pearson Education Ltd., Singapore.
8) Expanding the commercial value of PHP to the Enterprise (2005), available at http://www.intranetjournal.com/articles/200105/wp_05_24_01 a html
9) Felming, J., (1998), "Web navigation: designing the user experience", O' Reilly Publishers.
10) Getting started with Perl and CGI available at http://www.perl.com/cgi
11) Getting started with Microsoft FrontPage available at http://www. microsoft.com/library/en-us/frontpage.html
12) Getting started with JavaScript available at http://www.javaScript.com
13) Gilson, S., (2002), "Developing ColdFusion MX Aplications with CFML", available at http://www.macromedia.com/php/cioldfiusion/documentation/fmx-dev-cf-apps.pdf.
14) HTML and XHTML Tutorial available at http://www.webr eference.com/xm/reference/xhtml.html
15) Introduction to VBScript available at http://www.Vbxml.com/xhtml/articles/xhtml-tables
16) Introduction to XHTML available at http://www.w3schools.com/xhtml/default.asp
17) JavaScript tutorial repository available at http://www.javaScripts search.com
18) Jeske, D., (2005), "Clearsilve compared: VS PHP, ASP, JSP" available at http://www.clearsilver.net
19) Krill, Paul (2008-06-23)."JavaScript creator ponders past, future". InfoWorld.http://www.infoworld.com/article/08/06/23/eich-javascript-interview_1.html. Retrieved 2009-05-19.
20) Macromedia Dreamweaver MX repository available at http://www.macromedia.com/software/dreamwever.
21) Macromedia Flash MX repository available at http;//www.macromeida.com/ software/flash
22) National Institute for Standard and Technology, (1998), "WebMetric Tools", available at http://zing.ncsl.nist.gov/webnet, 1998.
23) Nielsen, J., (1999), "Designing Web usability: the practice of simplicity", New Riders Publishing.
24) Nielsen, J., and Mack, R., (1994), 3rd Edition, "Usability Inspective Methods", Wiley Publishers. Ovum Evaluates: Web Development Tools available at http://www.dpu.se/o vuweb-e.html-32k
25) PHP repository available at http://www.php.net.
26) Pandey, R.K. and Burnett, M.M. (1993) "Is it Easier to Write Matrix Manipulation Programs Visually or Textually? An Empirical Study", in IEEE Symposium on Visual Languages, Bergen, Norway, pp. 344-351.
27) Petreley, N., (2001), "Server-side HTML hell", available at http://www.itworld.com/AppDev/4072/LWD01053/penguin4/-46k
28) Prechelt, L., (2000), An Empirical Evaluation of C. C++, Java, Perl, Pytton, Rexx and Tel. Journal of visual languages and computing, IEEE computer society press, Caliphonia, U.S.A.
29) Resenfeld, L., and Morville, P., (1998), "Information architecture for the World Wide Web", O' Reilly Publisher.The World Wide Web consortium, http://www.w3c.org/march 2000.
30) "Rice's Theorem". *The Perl Review***4** (3): 23–29. Summer 2008.and "Perl is Undecidable". *The Perl Review***5** (0): 7–11. Fall 2008., which is available online at Kegler, Jeffrey. "Perl and Undecidability". http://www.jeffreykegler.com/Home/perl-and-undecidability
31) Tim Berners-Lee (2001) "Design Issues" Available online en.wikipedai.org/wiki/HTML
32) Waites, N., and Knott, G., (1998): COMPUTING, 3rd Edition, Business Education Publishers Ltd, Sunderland, U.K.
33) Ward-Dutton, N., (2000), "Evaluation of Web based Information Technology" available at http://www.dpu.se/neilweb-e.html
34) Web building tutorials repository available at http://www.w3schools.com
35) WebCriteria (2000), Web-based : "Comparative Evaluation of a Website", available at http://www.webcriteria.com
36) XML repository available at http://www.xml.com

Missing Value Estimation In Microarray Data Using Fuzzy Clustering and Semantic Similarity

Mohammad Mehdi Pourhashem[1], Manouchehr Kelarestaghi[2], Mir Mohsen Pedram[3]

GJCST Classification
H.3.3, I.5.3

***Abstract*-Gene expression profiling plays an important role in a broad range of areas in biology. Microarray data often contains multiple missing expression values, which can significantly affect subsequent analysis. In this paper, a new method based on fuzzy clustering and genes semantic similarity is proposed to estimate missing values in microarray data. In the proposed method, microarray data are clustered based on genes semantic similarity and their expression values and missing values are imputed with values generated from cluster centers. Genes similarity in clustering process determine with their semantic similarity obtained from gene ontology as well as their expression values. The experimental results indicate that the proposed method outperforms other methods in terms of Root Mean Square error.**

Keywords-microarray, missing value estimation, fuzzy clustering, semantic similarity

I. Introduction

Microarray is a technology for the monitoring of thousands of gene expression levels simultaneously [1]. Data from microarray experiments are usually in the form of large matrices of expression levels of genes (rows) under different experimental conditions (columns). For a number of reasons, microarray data sets frequently contain some missing values; typical reasons include insufficient resolution, image corruption, spotting or scratches on the slide, dust or hybridization failures [2]. Therefore missing value estimation is essential as a preprocessing step to obtain proper results from microarray data analysis. There are several approaches to deal with missing values. The first approach is repeating the experiment [3], which is expensive and time consuming. The second approach is ignoring objects containing missing values [4], that usually loses too much useful information and may bias the results if the remaining cases are unrepresentative of the entire sample.The third approach is estimating the missing values, which can be subdivided into two groups. The first group doesn‘t consider the correlation structure among the genes. These methods substitute the missing values by a global constant such as 0 [4], or by the average of the available values for that gene [5]. Both of these methods distort relationships among variables. The second groups consider the correlation structure. In fact the estimating procedure consists of two steps: in the first step similar genes to the gene with missing value, are selected and in the second step the missing values are predicted using observed values of selected genes, for example the widely used weighted K-nearest neighbor imputation (KNNimpute), reconstructs the missing values using a weighted average of K most similar genes [6]. These methods have better performance than simple methods such as substituting missing values by a constant or by row average, but their drawback is that estimation ability of them depends on K parameter (number of gene neighbor used to estimate missing value). There is no theoretical way, however, to determine this parameter appropriately and should be specified by user. In [2, 7] cluster-based algorithms have been proposed to deal with missing values which don‘t need user to determine parameters [8].A limitation of the methods mentioned above, is that they use no external information but the estimation is based solely on the expression data. In [8] a method based on Fuzzy C-means clustering algorithm (FCM) and gene ontology have been proposed to avoid the problems of those methods. This method (FCMGOimpute) uses information of gene ontology as external information, furthermore microarray data. There‘s a prospect that similar genes have close expression levels. In FCMGOimpute method two genes will be similar if they have the same annotations. This similarity measure is not good enough.In this paper, we propose a new missing value estimation method based on Fuzzy C-means clustering algorithm (FCM) and genes semantic similarity to avoid the problems of previous methods and be more accurate in evaluate genes similarity.The structure of this paper is as follows: Section 2 describes FCMGOimpute method and the proposed method to enhance it. In Section 3, the experimental results are shown, and finally some discussions are given in Section 4.

II. Methods

The clustering aim is to decompose a given set of objects into subgroups or clusters based on similarity. Whereas each gene may be involved in more than one biological process, hard clustering methods which assign each gene to only one cluster can not ensure this characteristic of the genes [9]. We expect that single genes may belong to several clusters, and the clustering algorithm should handle incomplete data. With these requirements, FCM algorithm in [2] is a proper clustering algorithm. In the clustering process, we have used gene ontology annotation as external information to determine the semantic similarity of genes and acquire more biologically interpretable clusters.

About[1]- Computer Engineering Department, Islamic Azad University-Arak Branch, Arak, Iran,mmpourhashem@yahoo.com
About[2]- Computer Engineering Department, Tarbiat Moallem University, Karaj/ Tehran, Iran,kelarestaghi@tmu.ac.ir
About[3]- Computer Engineering Department, Tarbiat Moallem University, Karaj/ Tehran, Iran,pedram@tmu.ac.ir

1) *FCMGOimpute*

This method uses FCM clustering for cluster microarray data that is an incomplete data. Fuzzy clustering method allows one object to belong to several clusters. Each object belongs to a cluster with a membership degree between 0 and 1 [10].The data from microarray experiments is usually in the form of large matrices of expression levels of genes (rows) under different experimental conditions (columns). This matrix called*G* and a matrix *E*has been defined, where E_{ki} is equal to 0, if corresponding component in G (G_{ki}) is a missing value and equal to 1 otherwise.

Let $X = \{g_1, g_2, \dots, g_N\}$be the set of given genes of matrix G (g_i is i'th row of matrix G) and let c be the number of clusters. Then membership degree of data object g_k to cluster i is defined as u_{ik}, which holds the below constraints:

$$\sum_{k=1}^{N} \mathrm{u_{ik}} > 0 \qquad \forall \mathrm{i} \in \{1, \dots, \mathrm{c}\} \tag{1}$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad \forall k \in \{1, \dots, N\} \tag{2}$$

Fuzzy C-means clustering is based on minimization of the following objective function:

$$J(U, C) = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2} \tag{3}$$

where *m* is fuzziness parameter which is a real number greater than 1, and $d_{ik}{}^2$ is the Euclidean distance between data object g_k and cluster center i which is defined by:

$$d_{ik}^2 = \left||g_k - c_i|\right|^2 = \left(\frac{s}{e_k} \sum_{j=1}^{s} \left(g_{kj} {}_{-} c_{ij}\right)^2 e_{kj}\right)\left(1 - \frac{\sum_{t=1}^{N} u_{it}^{m} B_{kt}}{N}\right) \tag{4}$$

where s is the feature space dimension, $e_k = \sum_{j=1}^{s} e_{kj}$ and B_{kt} is defined based on gene ontology annotations of gene k and gene t, as follows:

$$B_{kt} = \begin{cases} 1 & \mathit{if}\ g_k\ \mathit{and}\ g_t\ \mathit{have\ the\ same\ annotation} \\ 0 & \mathit{otherwise} \end{cases}$$
$$, 1 \le t \le N \tag{5}$$

Therefore, the annotation of g_k is compared with annotation of all genes belonging to cluster i, more genes have the same annotation, more the distance shrink. Of course not all the genes have the same effects, therefore we multiply B_{kt} to the membership degree of gene g_t to cluster i; Consequently the genes which belong to cluster i with higher membership degree, have more effect [8].In case the g_k is an unknown gene, B_{kt} is equal to 0 for all t ($1 \le t \le N$), and consequently the second term of equation (4) is equal to 1 [8]. It leads to Euclidean distance which only consider gene expression levels and used in FCMimpute method.The algorithm minimizes the objective function shown in (3), by updating of the cluster centers and membership degrees, iteratively by Equation (6) and (7).

$$c_{ij} = \frac{\sum_{k=1}^{N} (u_{ik})^m e_{kj} g_{kj}}{\sum_{k=1}^{N} (u_{ik})^m e_{kj}} \tag{6}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}} \tag{7}$$

To determine the fuzziness parameter (m) and the number of clusters (c), some methods were proposed in [2].

2) Improving Similarity Criterion In FCMGOimpute By Using Genes Semantic Similarity

In FCMGOimpute two genes are similar, if they have same annotation, and they are dissimilar if theirannotations are different. According to this definition, similarity value will be 0 or 1. Since, similarity concept isn't crisp;we use semantic similarity as similarity criterion between genes,which will be a real value between 0 and 1.For example and further explain, suppose we have two genes that all of their annotation terms are equal except one. Based on similarity criterion in FCMGOimpute these genes are dissimilar and similarity measure will be 0, but their semantic similarity may be 0.9. While, similarity measure between these two genes must affect the clustering and estimation process by value of 0.9, not 0 To measure the semantic similarity between two genes, the first step is to establish the semantic similarity between their annotated GO terms in the ontology. One of the most widely used approaches is based on the information theory. Given a term *t*, the occurrence of *t*, *occur*(*t*) is defined as the number of times *t* occurs in the annotation database being analyzed. The frequency of the term *t*, *freq*(*t*) is the summation of occurrence of *t* and all its descendantsdefined as,

$$freq(t) = \sum_{t \in ancestors(t_i)} occur(t_i) \tag{8}$$

where *ancestors*(t_i) is the set of t_i's ancestors. This definition is based on the fact that if a gene product is annotated by a term, then it is also annotated by its parent terms. Therefore, given any term, we can estimate its probability of being directly or indirectly annotated by gene products in a corpus, which is defined as [11],

$$p(t) = \frac{freq(t)}{freq(t_{root})} \tag{9}$$

where t_{root} is the root term of the ontology that t belongs to. In GO, t_{root} could be Molecular Function (MF), Cellular Component (CC), or Biological Process (BP). Obviously, $p(MF) = p(CC) = p(BP) = 1$. Now, the information content of term t, *IC(t)* can be define as:

$$IC(t) = -\log[p(t)] \tag{10}$$

Given a pair of terms, $t_i$and t_j, their shared information content is defined as [11]:

$$share(t_i, t_j) = \max_{t \in S(t_i, t_j)} [IC(t)] \tag{11}$$

where $S(t_i,t_j)$ = $ancestors(t_i) \cap ancestors(t_j)$. Since $IC(t) \ge IC(ancestors(t))$, the maximum information content of their common ancestors should be the information carried by their least common ancestor [11].We will use Lin term semantic similarity [12] that is defined as:

$$TSim_{Lin}(t_i,t_j)=\frac{2\times share(t_i,t_j)}{IC(t_i)+IC(t_j)} \qquad (12)$$

We will use gene semantic similarity as: [11]

$$GSim(g_i,g_j)$$
$$=\frac{\sum_{t_i\epsilon T_i,t_j\epsilon T_j}TSim(t_i,t_j)}{|T_i|\cdot|T_j|-|T_i\cap T_j|^2+\sum_{t_i,t_j\epsilon T_i\cap T_j}TSim(t_i,t_j)} \qquad (13)$$

We modify the calculation of Euclidean distance in (4) as follows:

$$d_{ik}^2=||g_k-c_i||^2=$$

$$\left(\frac{s}{e_k}\sum_{j=1}^{s}(g_{kj}-c_{ij})^2e_{kj}\right)\left(1-\frac{\sum_{t=1}^{N}u_{it}^m GSim(g_k,g_t)}{N}\right) \qquad (14)$$

Calculation of cluster centers and membership degree is the same as (6) and (7).

3) Imputation of missing values

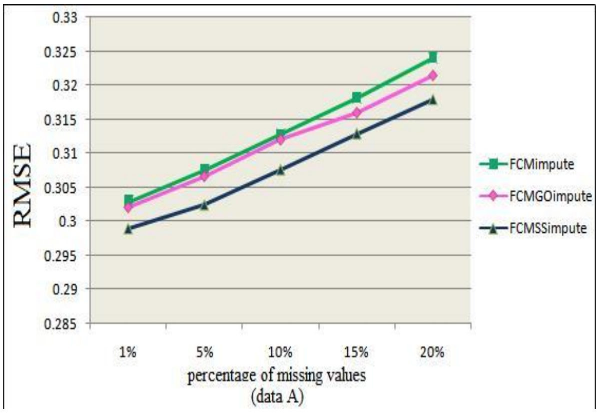
We utilize the clustering results to estimate the imputation of missing values in microarray data set. We impute missing values by making use of the weighted mean of the values of the corresponding attribute over all clusters. The weighting factors are the membership degrees u_{ik} of a gene g_k to the cluster c_i. The missing gene expression value g_{kj} is imputed by:

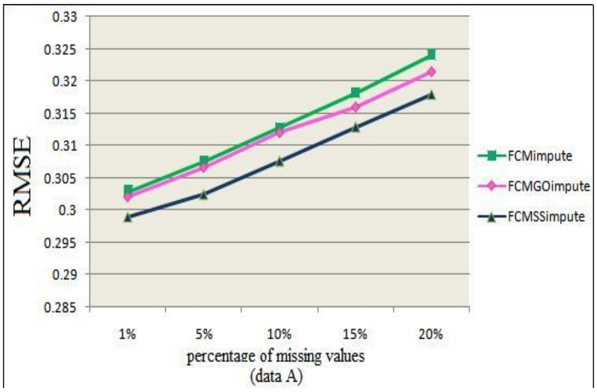
$$g_{kj}=\frac{\sum_{i=1}^{c}u_{ik}^m c_{ij}}{\sum_{i=1}^{c}u_{ik}^m} \qquad (15)$$

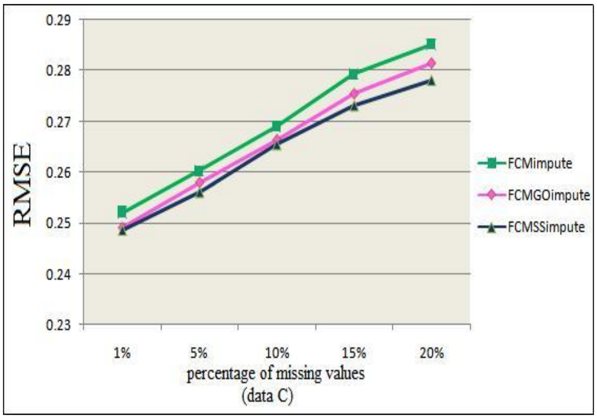
III. Experimental results

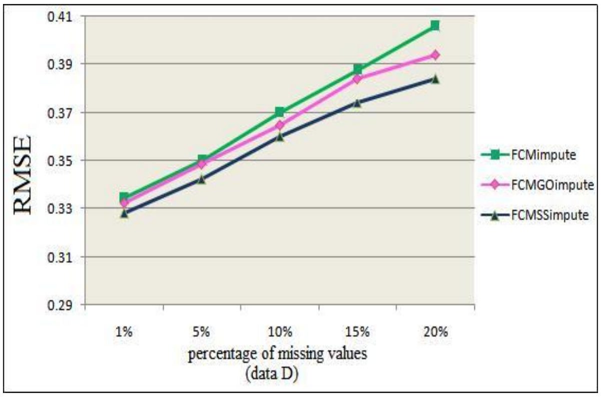
We compared our proposed method (FCMSSimpute) with the previously developed FCMimpute and FCMGOimpute methods by imputation of microarray data. Data set used in this work was selected from publically available microarray data. Five microarray were used: two microarray of yeast cells response to environmental changes, data A [13] and B [14], three microarray are time series of yeast, data C, D and E [15]. We collected GO annotation for the genes in thisdata set from [16] and necessary terms semantic similarity for compute genes semantic similarity from [17].Before applying the imputation algorithms, each data set was preprocessed for the evaluation by removing rows containing missing expression values, yielding complete‘ matrices. Between 1% and 20% of the data weredeleted at random to create test data sets. Each method was then usedto recover the introduced missing valuesfor each data set, and the estimated values were compared to those in the original data set. To compare the accuracy of different imputation methods, we used RMSE (Root Mean Squared Error) as evaluation metric:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(R_i-I_i)^2}{n}} \qquad (16)$$









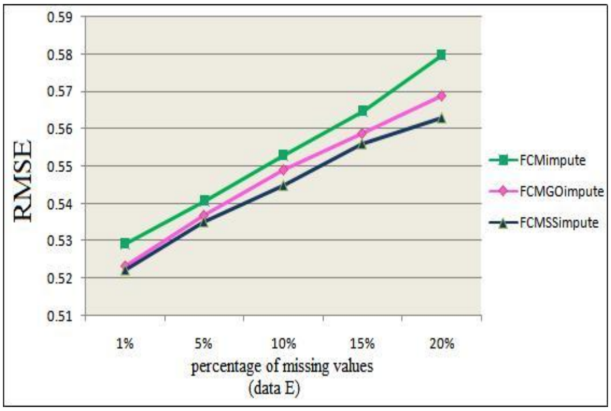


Figure 1. Comparison of the accuracy of FCMSSimpute, FCMGOimpute and FCMimpute methods for five data set over 1% and 20% data missing. The accuracies were evaluated by RMSE.

where, R is the real value, I is the imputed value, and n is the number of missing values.The FCMimpute considers the correlation structure amongst the genes, but doesn‘t use any useful external information such as gene ontology, and uses just microarray data for imputation process. As it can be seen from the results, the FCMimpute has a lower performance, compared to other methods. FCMGOimpute has better performance over FCMimpute because it uses gene ontology annotation as anexternal information.As it is clearly observed from the Figure 1, the proposed method (FCMSSimpute) outperforms others in terms of accuracy. The proposed method considers the correlation structure amongst the genes. Additionally, it uses gene ontology annotation as an external information, and genes semantic similarity to measure genes similarity, which is more accurate from what defined in FCMGOimpute. Therefore the accuracy of imputation based on a well defined similarity of genes, will increase.

IV. Conclusions

In this paper, we proposed a new and efficient method for estimating missing values in microarray data, based on the using of genes semantic similarity. We take advantage of the correlation structure of the data to estimate missing expression values by clustering, as well as using genes semantic similarity which improves the imputation accuracy.We have analyzed the performance of our method on fivemicroarray and compared the accuracy with FCMimpute and FCMGOimpute methods. We observed that our method outperforms other methods in terms of the RMSE.In this paper, we have used weighted majority vote to determine the similarity of a gene to a cluster. We have used semantic similarity for measure similarity between genes. To compute semantic similarity, we have used molecular function annotation of genes, but there exist alternatives to define semantic similarity by use other term semantic similarity measures. Also Biological Process annotations can be used in similarity computation.

V. References

1) Daxin Jiang, Jian Pei, Aidong Zhang, An Interactive Approach to Mining Gene Expression Data, IEEE Transactions On Knowledge And Data Engineering, vol. 17, no. 10, pp.1363-1378, October 2005.
2) J. Luo, T. Yang, Y. Wang, Missing Value Estimation For Microarray Data Based On Fuzzy C-means Clustering, in Proceedings of the Eighth International Conference on High-Performance Computing in Asia-Pacific Region, 2005.
3) A. J. Butte, J. Ye, Determining Significant Fold Differences in Gene Expression Analysis, Pac. Symp. Biocomput, vol. 6, pp. 6- 17, 2001.
4) A. A. Alizadeh and et al, Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling, Nature, vol. 403, pp. 503-511, 2000.
5) J. L. Schafer, J. W. Graham, Missing data: our view of the state of the art, Psychol. Methods, vol. 7, pp. 144- 177, 2002.
6) O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics, vol. 17, pp. 520- 525, 2001.
7) S. Zhang, J. Zhang, X. Zhu, Y. Qin, C. Zhang, Missing Value Imputation Based on Data Clustering, Transactions on Computational Science (TCOS) 1, pp. 128-138, 2008.
8) Azadeh Mohammadi, Mohammad Hossein Saraee, Estimating Missing Value in Microarray Data Using Fuzzy Clustering and Gene Ontology, IEEE International Conference on Bioinformatics and Biomedicine, pp. 382-385, 2008.
9) J. Shaik, M. Yeasin, Two-way Clustering using Fuzzy ASI for Knowledge Discovery in Microarrays, in Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2007.
10) J. Valente de Oliveira, W. Pedrycz, Advances in Fuzzy Clustering and its Applications: John Wiley & Sons, Ltd, 2007.
11) Zheng Chen, Jian Tang, Using Gene Ontology to Enhance Effectiveness of Similarity Measures for Microarray Data, BIBM, pp. 66-71, 2008.
12) Dekang Lin. An Information-Theoretic Definition of Similarity, In Proc. 15th International Conf. on Machine Learning, pp. 296-304, 1998.
13) http://homepages.nyu.edu/~rb133/, 28 June 2010.

14) http://titan.biotec.uiuc.edu/rox1/download.html, 28 June 2010
15) http://www-genome.stanford.edu/mec1/data.shtml 28 june 2010
16) Saccharomyces Genome Database, http://db.yeastgenome.org/cgi-bin/batchDownload, 20 June 2010.
17) http://bioinformatics.clemson.edu/G-SESAME/Program/GOCompareMultiple1.php.

Generalized Hadamard Matrices from Generalized Orthogonal Matrix

Sudha Singh[1],M. K. Singh[2], D. K. Singh[3]

GJCST Classification
G.1.3

***Abstract*-A new generalization of matrix orthogonality is introduced. It is shown thatfrom generalized orthogonal matrices some known as well as a few new complex H- matrices with circulant blocks can be obtained. The orders of new complex H-matrix are 26, 36, 50 and 82.**

IndexTerms- Circulant matrix, Hadamard matrix, Generalization of Hadamard matrix, Quaternion, Associative algebra of matrices, generalized orthogonal matrix.

I. Introduction

First we recall the following definitions:
Circulant Matrix: It is an $\boldsymbol{n\times n}$ matrix of the form

$$\begin{pmatrix} a_1 & a_2 & a_3 \ldots\ldots a_n \\ a_n & a_1 & a_2 \ldots\ldots a_{n-1} \\ a_{n-1} & a_n & a_1 \ldots\ldots a_{n-2} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ a_2 & a_3 & a_4 \ldots\ldots a_1 \end{pmatrix}$$

1 4 3 2 2 1 1 1 2 1 3 2 1

a a a a a a
a a a a a a a a
a a n n n n n n

which is denoted as $Circ(a_1 a_2 a_3 \ldots\ldots a_n)$.

Hadamard matrix (or an H-matrix): It is an $n\times n$ matrix H with entries +1, -1 such that $HH^T = nI_n$, where I_n is the $\boldsymbol{n\times n}$ identity matrix.

Complex H-matrix: It is an $\boldsymbol{n\times n}$ matrix $H=[H_{ij}]$, where H_{ij} are complex numbers with $|H_{ij}|=1$ for i, j = 1, 2, . . . , n, satisfying $HH^* = nI$, where I is the identity matrix and H^* denotes the Hermitian transpose[9] of H. A complex H-matrix is called dephased if elements of its first row and column are 1.

Butson H-matrix: It is an $\boldsymbol{n\times n}$ complex Hadamard matrix with elements belonging to theset of $\boldsymbol{m}^{th}$ roots of 1 and is denoted as $BH(m,n)$.

Unimodular complex H-matrix: It is an $\boldsymbol{n\times n}$ complex H-matrix whose elements are of the form $\boldsymbol{EXP(i\theta)}$. An _

m-parameteraffine complex Hadamard family(or orbit) H(R) stemming from a dephased $\boldsymbol{n\times n}$ complex Hadamard matrix H is the set of matrices A satisfying $\boldsymbol{AA^* = nI}$, associated with an m-dimensional subspace R of a space of all real $n\times n$ matrices with zeros in the first row and column,

Weighing matrix W(n,w): A W(n,w) of order *n* and weight *w* is an $n\times\boldsymbol{n}$ (0,1, − 1)-matrix such that $\boldsymbol{WW^T = wI}$, where $\boldsymbol{w}$ is a positive integer.

Conference matrix: It is a weighing matrix $W(\boldsymbol{n},n-\mathbf{1})$ with0 occurring only on the diagonal.

Quaternion: A number of the form q = a1 + bi + cj + dk, where $i^2 = j^2 = k^2 = -1$, k = ij = -ji, a, b, c, d are real numbers, is called a quaternion or a hypercomplex number. q reduces to a complex number when c = d = 0 and to a real number when b = c = d = 0. If 1, i, j, k are taken as

2×2 matrices $\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} \mathbf{0} & 1 \\ -1 & 0 \end{pmatrix}$, $\begin{pmatrix} \mathbf{-i} & \mathbf{0} \\ 0 & i \end{pmatrix}$ **and** $\begin{pmatrix} \mathbf{0} & \mathbf{i} \\ i & 0 \end{pmatrix}$

respectively we get two dimensional complex matrix representation of the quaternion q = $\begin{pmatrix} a - ci & b + di \\ -b + di & a + ci \end{pmatrix}$.

ci a

Replacing 1 by $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and i by $\begin{pmatrix} \mathbf{0} & \mathbf{1} \\ -1 & 0 \end{pmatrix}$ in the matrices 1, i, j, k
we get four dimensional real matrix

representation of the quaternion q = $\begin{pmatrix} \mathbf{a} & -\mathbf{c} & \mathbf{b} & d \\ \mathbf{c} & \mathbf{a} & \mathbf{-d} & \mathbf{b} \\ -b & d & a & c \\ -d & -b & -c & a \end{pmatrix}$.

The Hermitian conjugate of a quaternion q = a1 + ib + jc + kd is $\overline{\boldsymbol{q}}$ = a1 - ib - jc – kd and the modulus of q is $|q| = \sqrt{a^2+b^2+c^2+d^2}$.

Associative algebra of matrices: Let R be a ring of complex numbers and A be a vector space of vxv matrices over R with basis matrices I =A_0,A_1,. . .A_m. A is called an associative algebra with unity I if they satisfy

A_iA_j = $\sum_{k=0}^{m} p_{ij}^{k} A_k$, where p^k_{ij} are in general complex numbers (1)

About[1]-PG Department of Computer Secience and Engg, Bengal College of Engg and Technology, Durgapur-713212(West Bengal),India;sudha_2k6@yahoo.com

About[2]- PG Department of Mathematics, Ranchi University, Ranchi-834008(Jharkhand), India;mithileshkumarsingh@gmail.com

About[3]- Department of Electronics and Communications BIT Sindri, Dhanbad-(828123) (Jharkhand), India, dksingh_bit@yahoo.com.

In what follows we assume that the elements of A_i are only 0, 1 and –1 and p^k_{ij} are integers called multiplication coefficients of the algebra.

Example 1 Algebra of quaternions spanned by the matrices 1, i, j, k of order 2 or 4 given in 1.4.

Example 2 Algebra of circulant matrices spanned by $A_i = w^i = [\mathrm{circ}(0,1, 0,. . .0)]^i$,
i= 1,2,…m satisfying $A_iA_j= A_{i+j}$, where i+j is the addition mod m. We also consider
the algebra spanned by the direct product of circulant matrices of different orders viz.
$w_s^i \times w_t^{j,}$ where w_s= circ(01,0,. . .0) of order s.

Example 3 Bose-Mesner algebra spanned by (0, 1) symmetric commuting matrices A_i satisfying
$A_0+A_1 + ...+A_m = J_v$ (all 1 matrix) and (1), where $A_0= I_v$,and p^k_{ij} are nonnegative integers.
(A_0, A_1 ...A_m) defines an m-class association scheme(or m-AS) with parameters p^k_{ij} .
A 2-AS is also called strongly regular graph and its parameters satisfy
$p^0 ii = n_i$, i=1, 2, and p^k_{ij} . satisfy
$p^k_{ij} = p^k_{ji}$, $p^j_{i0}= \delta_{ij}$, n_1+n_2= v-1, and $p^i_{j1} + p^i_{j2}= n_j- \delta ij$.,
i,j=1,2, where δij=0 for i≠j and
δij=1 for i=j. (see Raghavarao[4]).

Generalized orthogonal matrix (GOM): Let A be an $m \times n$ matrix whose entries are the element of an associative algebra of matrices over a ring of complex numbers. The conjugate of an element a = $\sum_{g\in G} \alpha_i A_i \in A$ will be denoted by $\bar{a} = \sum_{g\in G} \overline{\alpha_i} A_i^T$, where $\overline{\alpha_i}$ is the complex conjugate of α_i and T stands for transpose.

A will be called a generalized orthogonal matrix if the dot product of any two rows

$$R_i.R_j = (a_{i1}, a_{i2} ..., a_{in})(b_{j1}, b_{j2} ..., b_{jn})$$

$$= \sum_{k=1}^{n} a_{ik}\overline{b_{jk}} =$$

$$\begin{cases} \lambda J, \ if \ i \neq j \\ \lambda_0 I + \lambda_1 \sum_{i=1}^{m} \mathrm{A_i} \quad \text{if} \quad \mathrm{i = j,} \end{cases}$$

where λ, λ_0, λ_1 are integers independent of i and j. Here λ, λ_0, λ_1 will be called parameters of orthogonal matrix A.

The purpose of this paper is to show that notion of generalized orthogonal matrix provides a general frame-work for constructing several classical real H-matrices of Paley[3] Williamson[7] and Ito [1] as well as some new Butson H-matrices and GDG H-matrices through special methods or computer search.We also identify some Butson H-matrices which admit non-Dita-type affine complex Hadamard family(or orbit)(vide sz¨oll˝osi [5]) Such matrices are recently being used in quantum information theory and quantum tomography.Notations: The circulant matrix circ(0,1,0,…,0)will be denoted as $\mathrm{w_n}$. The direct product of $\mathrm{w_m}, \mathrm{w_n}$ will be denoted as $w_m x w_n$.

II. Construction of complex H-matrices from generalized orthogonal matrices

A. *Construction of some H-matrices with circulant blocks*

Construction of certain Paley type-I H-matrices [for definition see page 12,chapter 2 of 10]

Theorem I : Let $p = 4t - 1$ be a prime. If $(d_1, d_2, d_3,..., d_k) \bmod p$ be a difference set[11,10], then GO-matrix $A = [w_p^{d_1} + w_p^{d_2} + ... + w_p^{d_k}]$, where $w_p = Circ(0,1,0,0,...,0)_p$ gives the core of a H-matrix of order 4t, if we replace 0 by -1 in A.

Construction of H-matrices of Williamson's form[10]

Williamson H-matrix of order $4(2m+1)$ is itself a 1×1 generalized orthogonal matrix
H = 1 x A + i x B + j x C + k x D, where 1, i, j, k are 4x4 matrix representation of basic quaternions,

$$1 = I_4, \mathrm{i} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \mathrm{j} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \mathrm{k} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \text{and} \quad \mathrm{A,}$$

B, C, D are (+1,-1) suitable linear combinations of (0,1)-circulant matrices $W_1, W_2, W_3,..., W_m$ of order n, that span a Bose-mesner algebra(see Raghavarao[4], for details of Bose-mesner algebra).

Construction of whiteman's type H–matrices of order (pq+1) where p and q are twin primes (vide whiteman[8])

Theorem II :Let p and q be two primes, q = p+2.
An 1x1 generalized orthogonal matrix

$$A = [(1 + w_p + w_p^2 + ... + w_p^{p-1}) \times I_q + (w_p \times w_q)^{d_1} + (w_p \times w_q)^{d_2} + ... + (w_p \times w_q)^{d_k}]$$

where

$$w_p = Circ(0,1,0,0,...,0)_p$$

$$w_q = Circ(0,1,0,0,...,0)_q$$

$$\text{and d satisfies} \, d^k \equiv 1(\mathrm{mod}\, p),$$

$$d^k \equiv 1(\mathrm{mod}\, q), \mathrm{k} = \frac{\mathrm{(p-1)(q-1)}}{2}$$

gives the core of H-matrix of order (pq+1), if we replace 0 by -1 everywhere in A.

Construction of an H-matrix of order 36

We consider on 3x7 rectangular generalized orthogonal matrix A=

$$\begin{pmatrix} \omega+\omega^4 & \omega^2+\omega^3 & 0 & 0 & I_5 & I_5 \\ I_5 & I_5 & \omega+\omega^4 & \omega^2+\omega^3 & I_5 & 0 \\ 0 & 0 & I_5 & I_5 & I_5 & \omega+\omega^4 \end{pmatrix}$$, where

$\omega = circ\ (01000),\ 0 = 5 \times 5$ null matrix and I_5 = unit matrix.

Then replacing 3 by 0 and 0 by -1 in A^TA, we get the core of a H-matrix (see Horadam[10] for definition and details of core) of order 36.

B. H-matrices from generalized orthogonal matrices arising from BIBDs (see Hall [12] for BIBDs)

Theorem III: Existence of a BIBD with parameters $v = 2n^2 - n, b = 4n^2 - 1, r = 2n+1, k = n, \lambda = 1$ implies the existence of an H- matrix of order $4n^2$.

Method of construction: Let N be the incidence matrix of BIBD with parameters mentioned in theorem III. N^tN is a $b \times b$ square matrix. Let A be the (1,-1) matrix obtained from N^tN by replacing diagonal entries by -1, 1 by 1 and 0 by -1. Then A is a 1×1 generalized orthogonal matrix and $\begin{pmatrix} -1 & e \\ e^t & A \end{pmatrix}$ is a H-matrix of order $4n^2$ where e is $1 \times (4n^2 - 1)$ matrix of 1's, e^t is the transpose of e.

Example 4: We consider the BIBD
Parameters: v=6, b=15, r=5, k=2, λ=1.

Let N be the incidence matrix of BIBD with given parameters.
Its dual N' is

$$\begin{pmatrix} 000011 \\ 110000 \\ 101000 \\ 100100 \\ 100010 \\ 100001 \\ 011000 \\ 010100 \\ 010010 \\ 010001 \\ 001100 \\ 001010 \\ 001001 \\ 000110 \\ 000101 \end{pmatrix}$$

The product of N and N' is

$$\begin{pmatrix} 2111111110\ 00000 \\ 1211110001\ 11000 \\ 1121101001\ 00110 \\ 1112100100\ 10101 \\ 1111200010\ 01011 \\ 1100021111\ 11000 \\ 1010012111\ 00110 \\ 1001011210\ 10101 \\ 1000111120\ 01011 \\ 0110011002\ 11110 \\ 0101010101\ 21101 \\ 0100110011\ 12011 \\ 0011001101\ 10211 \\ 0010101011\ 01121 \\ 0001100110\ 11112 \end{pmatrix}$$

$= 2A_0 + 1A_1 + 0A_2$

where $I = A_0, A_1, A_2$ span Bose-Mesner algebra. From NN' we can obtain a 1×1 generalized orthogonal matrix A by replacing 2 by 0 and 0 by -1. Adjoining a row of all 1's and a column of all 1's we get the following 16×16 H-matrix

$$\begin{pmatrix}
-1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\
1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 \\
1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1
\end{pmatrix}$$

This matrix attains an affine orbit by lemma 3.4 (see SZOLLOSI [5])

Example 5: We consider the BIBD
Parameters: v =15, b=35, r=7, k=3, λ=1.

Let N be the incidence matrix of BIBD with given parameters. Its dual N' is

Fig 1: MATRIX-1

The product of N and N' is

Fig 2: MATRIX-2

$$=3A_0+1A_1+0A_2$$

where $I=A_0, A_1, A_2$ span Bose-Mesner algebra. From NN' we can obtain a 1×1 generalized orthogonal matrix A by replacing 3 by 0 and 0 by -1. Adjoining a row of all 1‘s and a column of all 1‘s we get the following 36×36 H-matrix

Fig 3: MATRIX-3

This matrix attains an affine orbit by lemma 3.4 (see SZOLLOSI [5]).

C. Construction of some new Butson H-matrices

Example 6: Butson H-matrices can be constructed from the following circulant representation of some generalized orthogonal matrices:

(i) $A_5=[w+w^4 \quad w^2+w^3]$,

where $w=w_5=circ(01000)$

(ii) $A_{13}=[w+w^3+w^9 \quad w^2+w^5+w^6]$, where

$w=w_{13}=circ(010\ldots0)$ of order13

(iii) $A_5=[circ(1+w^4, w, 0, 0, 1) \ \text{circ}(1+w^3, 0, w^2, 1, 0)]$,

$where \mathrm{w}=\mathrm{w}_5=circ(01000)$

(iv) $A_{41}=[w+w^{37}+w^{16}+w^{18}+w^{10} \quad w^8+w^9+w^5+w^{21}+w^{39}]$.

where $w=circ(01\ldots0)$ of order 41.

Method of Construction: Let A be any of the matrices above in (i), (ii),(iii) or (iv).

a) Obtain the symmetric square matrix A^tA.

b) **In A^tA replacing diagonal element by 1, 0 by –i and 1 by i, we get Butson** H-matrices $BH(4,2n)$ for 2n= 10, 26, 50 and 82.

c) In A^tA replacing diagonal elements by 0, we get a conference matrix.

Remark 1 The matrices of above orders constructed from circulant matrices of order 5, 13 and 41 appears to be different from those arising from well-known constructions from Galois fields of order 25, 49 and 81.

Remark 2: Since Hadamard matrices obtained in the above theorem are derivable from conference matrices, each matrix A is non Dita-type and admits an affine family of complex Hadamard matrices of at least one parameter which contains A (vide sz¨oll′′osi‘s theorems 4.1 and 4.2 in [5]).

Remark 3: In the recent catalogue [6] only Dit¸˜a-type matrices were considered in dimensions N = 10 and 14. Sz¨oll′′osi [5] presents non Dita-type matrix of order 10. In view of Theorem 4.1 and of 4.2 of sz¨oll′′osi we can now present new parametric families of non Dita-type complex Hadamard matrices of order 26, 50, 82.

(1) $BH(4,26)$

A Butson H-matrix of order 26 obtained by the above method is :

Fig 4: MATRIX-4

(2) $BH(4,50)$

A Butson H-matrix of order 50 obtained by the above method is :

Fig 5: MATRIX-5

D. Some Butson H-matrices $BH(m,n)$ for $m=3,6$.

Following Butson H-matrices are obtained from suitable generalized orthogonal matrices

(i) $BH(3,6)$: $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & & & & & \\ 1 & & A & & & \\ 1 & & & & & \\ 1 & & & & & \\ 1 & & & & & \end{pmatrix}$ where the core A is

given by $A=[I_5+w(w_5+w_5^4)+w^2(w_5^2+w_5^3)]$,

where $w_5=circ(01000)$ and w is an imaginary cube root of unity.

(ii) BH(3,9)

$$\begin{pmatrix} 1 & w & w & w & w & w^2 & w^2 & w^2 & w^2 \\ w & 1 & w & w^2 & w^2 & w & w & w^2 & w^2 \\ w & w & 1 & w^2 & w^2 & w^2 & w^2 & w & w \\ w & w^2 & w^2 & 1 & w & w & w^2 & w & w^2 \\ w & w^2 & w^2 & w & 1 & w^2 & w & w^2 & w \\ w^2 & w & w^2 & w & w^2 & 1 & w & w & w^2 \\ w^2 & w & w^2 & w^2 & w & w & 1 & w^2 & w \\ w^2 & w^2 & w & w & w^2 & w & w^2 & 1 & w \\ w^2 & w^2 & w & w^2 & w & w^2 & w & w & 1 \end{pmatrix}$$

(iii) BH(6, 7)

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -w^2 & -1 & -1 & w & -w & -w \\ 1 & -1 & -w^2 & -1 & -w & w & -w \\ 1 & -1 & -1 & -w^2 & -w & -w & w \\ 1 & w & -w & -w & -w^2 & -1 & -1 \\ 1 & -w & w & -w & -1 & -w^2 & -1 \\ 1 & -w & -w & w & -1 & -1 & -w^2 \end{pmatrix}$$

III. Conclusion

Butson H-matrices are constructed from generalized orthogonal matrices by replacement or minor changes. During constructions we get new complex H-matrices of orders 26, 36, 50 and 82, which is not equivalent to existing complex Hadamard matrices of same order. We hope that in future generalized orthogonal matrices will provide insights to construct more matrices of combinatorial and practical interests.

IV. References

1) Ito N., On Hadamard groups III, Kyushu J. Math. 51(1997) 1-11.
2) Jafarkhani, H. A quasi-orthogonal Space-Time Block Code. IEEE Trans. Communications,49, 1-4.(2001).
3) Paley, R.E. A.C.. On orthogonal matrices. J. Math. Phys., 12:311--320, 1933.
4) Raghavarao D., Constructions and Combinatorial Problems in Design of Experiments. Dover, New York, 1988.
5) Szollosi F., Parametrizing complex Hadamard matrices" European journal of combinatorics , V29 No5, (2008) 1219-1234.
6) Tadej W. and K. Zyckkowski, A concise guide to complex Hadamard matrices,
Open Sys. & Information Dyn. 13(2006) 133-177.
7) Williamson J.. Hadamard's determinant theorem and the sum of four squares. Duke Math. J., 11:65--81, 1944.
Whiteman, A. I. A family of difference sets" Illnois J. Math. 6(1962), 107-121.
8) Whiteman, A. I. A family of difference sets" Illnois J. Math. 6(1962), 107-121.
9) Meyer, C. D. Matrix Analysis and Applied Linear Algebra. Philadelphia, PA: SIAM, 2000.
10) Horadam K. J, Hadamard Matrices and their Applications", Princeton University Press, 2007.
11) Weisstein, Eric W. "Difference Set." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/DifferenceSet.html
12) Hall M., Combinatorial Theory" , Wiley-Interscience July 2, 1998

Fig 1: Matrix-1

$$\begin{pmatrix} 0100110000\ 00000 \\ 1010001000\ 00000 \\ 0101000100\ 00000 \\ 0010100010\ 00000 \\ 1001000001\ 00000 \\ 0011010000\ 00000 \\ 0001101000\ 00000 \\ 1000100100\ 00000 \\ 1100000010\ 00000 \\ 0110000001\ 00000 \\ 0000001001\ 10000 \\ 0000010100\ 01000 \\ 0000001010\ 00100 \\ 0000000101\ 00010 \\ 0000010010\ 00001 \\ 0000000110\ 10000 \\ 0000000011\ 01000 \\ 0000010001\ 00100 \\ 0000011000\ 00010 \\ 0000001100\ 00001 \\ 1000010000\ 10000 \\ 0100001000\ 01000 \\ 0010000100\ 00100 \\ 0001000010\ 00010 \\ 0000100001\ 00001 \\ 1000000000\ 01001 \\ 0100000000\ 10100 \\ 0010000000\ 01010 \\ 0001000000\ 00101 \\ 0000100000\ 10010 \\ 1000000000\ 00110 \\ 0100000000\ 00011 \\ 0010000000\ 10001 \\ 0001000000\ 11000 \\ 0000100000\ 01100 \end{pmatrix}$$

Fig 2: Matrix-2

$$\begin{pmatrix}
3011011111 & 0100100110 & 1100101001 & 01001 \\
0301111111 & 1010000011 & 1110010100 & 10100 \\
1030111111 & 0101010001 & 0111001010 & 01010 \\
1103011111 & 0010111000 & 0011100101 & 00101 \\
0110311111 & 1001001100 & 1001110010 & 10010 \\
1111131001 & 0100100110 & 1011000110 & 00110 \\
1111113100 & 1010000011 & 0101100011 & 00011 \\
1111101310 & 0101010001 & 1010110001 & 10001 \\
1111100131 & 0010111000 & 1101011000 & 11000 \\
1111110013 & 1001001100 & 0110101100 & 01100 \\
0100101001 & 3011011111 & 1100101001 & 00110 \\
1010010100 & 0301111111 & 1110010100 & 00011 \\
0101001010 & 1030111111 & 0111001010 & 10001 \\
0010100101 & 1103011111 & 0011100101 & 11000 \\
1001010010 & 0110311111 & 1001110010 & 01100 \\
0011000110 & 1111131001 & 1011001001 & 00110 \\
0001100011 & 1111113100 & 0101110100 & 00011 \\
1000110001 & 1111101310 & 1010101010 & 10001 \\
1100011000 & 1111100131 & 1101000101 & 11000 \\
0110001100 & 1111110013 & 0110110010 & 01100 \\
1100110110 & 1100110110 & 3000011001 & 10110 \\
1110001011 & 1110001011 & 0300011100 & 01011 \\
0111010101 & 0111010101 & 0030001110 & 10101 \\
0011111010 & 0011111010 & 0003000111 & 11010 \\
1001101101 & 1001101101 & 0000310011 & 01101 \\
0100100110 & 0100101001 & 1100130110 & 11111 \\
1010000011 & 1010010100 & 1110003011 & 11111 \\
0101010001 & 0101001010 & 0111010301 & 11111 \\
0010111000 & 0010100101 & 0011111030 & 11111 \\
1001001100 & 1001010010 & 1001101103 & 11111 \\
0100100110 & 0011000110 & 1011011111 & 31001 \\
1010000011 & 0001100011 & 0101111111 & 13100 \\
0101010001 & 1000110001 & 1010111111 & 01310 \\
0010111000 & 1100011000 & 1101011111 & 00131 \\
1001001100 & 0110001100 & 0110111111 & 10013
\end{pmatrix}$$

Fig 3: Matrix-3

$$\begin{pmatrix}
1 \; 1 \; 1 \; 11 \; 1 \; 11111 \; 1 \; 1 \; 1 \; 11 \; 1 \; 111 \; 111 \; 1 \; 11 \; 11 \; 1 \; 11 \; 11 \; 1 \; 11 \\
1-1-111-111111 \; -11-1-11-1-111-111-1-11-11-1-11-11-1-11 \\
1-1-1-111111111 \; -11-1-1-1-1-111111 \; -1-11-11-1-11-11-1-1 \\
11-1-1-1111111 \; -11-11-11-1-1-11-1111-1-11-11-1-11-11-1 \\
111-1-1-111111 \; -1-11-1111-1-1-1-1-1111-1-11-11-1-11-11 \\
1-111-1-1111111 \; -1-11-1-111-1-11-1-1111-1-11-11-1-11-1 \\
111111 \; -11-1-11-11-1-11-1-111-11-111-1-1-111-1-1-111-1 \\
1111111 \; -11-1-11-11-1-1-1-1-111-11-111-1-1-111-1-1-111 \\
111111 \; -11-11-1-11-11-11-1-1-111-11-111-1-1-111-1-1-11 \\
111111 \; -1-11-11-1-11-1111-1-1-111-11-111-1-1-111-1-1-1 \\
1111111 \; -1-11-11-1-11-1-111-1-1-111-11-111-1-1-111-1-1 \\
1-11-1-11-11-1-11-1-111-11111111 \; -1-11-11-1-11-1-111-1 \\
11-11-1-11-11-1-1-1-1-1111111111 \; 1-1-11-11-1-1-1-1-111 \\
1-11-11-1-11-11-11-1-1-1111111 \; -1111-1-11-11-11-1-1-11 \\
1-1-11-11-1-11-1111-1-1-111111 \; -1-1111-1-11-1111-1-1-1 \\
11-1-11-11-1-11-1-111-1-1111111 \; -1-1111-1-11-1-111-1-1 \\
1-1-111-1-1-111-111111 \; -11-1-111-111-1-11-1-11-1-111-1 \\
1-1-1-111-1-1-111111111 \; -11-1-1-11-1111-11-1-1-1-1-111 \\
11-1-1-111-1-1-1111111 \; -11-11-11-11-11-11-11-11-1-1-11 \\
111-1-1-111-1-1-111111 \; -1-11-1111-11-1-1-11-1111-1-1-1 \\
1-111-1-1-111-1-1111111 \; -1-11-1-111-111-1-11-1-111-1-1 \\
111-1-111-111-111-1-111-111-1-1-1-1-1-111-1-111-111-1 \\
1111-1-1-11-111111 \; -1-1-11-111-1-1-1-1-1111-1-1-11-111 \\
1-1111-11-11-11-1111-11-11-11-1-1-1-1-1-1111-11-11-11 \\
1-1-111111 \; -11-1-1-111111 \; -11-1-1-1-1-1-1-1-111111 \; -11-1 \\
11-1-111-111-111-1-111-111-11-1-1-1-1-11-1-111-111-11 \\
1-11-1-11-1-111-1-11-1-11-11-1-1111-1-11-1-111-111111 \\
11-11-1-1-1-1-1111-11-1-11-11-1-1111-1-1-1-1-11111111 \\
1-11-11-11-1-1-11-11-11-1-11-11-1-1111-11-1-1-1111111 \\
1-1-11-1111-1-1-1-1-11-11-1-11-11-1-111111 \; -1-1-111111 \\
11-1-11-1-111-1-11-1-11-11-1-11-11-1-111-111-1-111111 \\
1-11-1-11-1-111-1-1-111-1-1-111-11-111-111111 \; -11-1-11 \\
11-11-1-1-1-1-111-1-1-111-1-1-111-11-111111111 \; -11-1-1 \\
1-11-11-11-1-1-111-1-1-111-1-1-111-11-1111111 \; -11-11-1 \\
1-1-11-1111-1-1-111-1-1-111-1-1-111-11-111111 \; -1-11-11 \\
11-1-11-1-111-1-1-111-1-1-111-1-1-111-11111111 \; -1-11-1
\end{pmatrix}$$

Fig 4: Matrix-4

$$\begin{pmatrix}
i-11-1-11111 -1-11-1-1111111 -1-111-11 \\
-1i-11-1-11111 -1-111-1111111 -1-111-1 \\
1-1i-11-1-11111 -1-1-11-1111111 -1-111 \\
-11-1i-11-1-11111 -11-11-1111111 -1-11 \\
-1-11-1i-11-1-1111111 -11-1111111 -1-1 \\
1-1-11-1i-11-1-1111 -111-11-1111111 -1 \\
11-1-11-1i-11-1-111-1-111-11-1111111 \\
111-1-11-1i-11-1-111-1-111-11-111111 \\
1111 -1-11-1i-11-1-111-1-111-11-11111 \\
-11111 -1-11-1i-11-1111-1-111-11-1111 \\
-1-11111 -1-11-1i-111111 -1-111-11-111 \\
1-1-11111 -1-11-1i-111111 -1-111-11-11 \\
-11-1-11111 -1-11-1i111111 -1-111-11-1 \\
-11-111-1-1111111 i1-111-1-1-1-111-11 \\
1-11-111-1-1111111 i1-111-1-1-1-111-1 \\
11-11-111-1-11111 -11i1-111-1-1-1-111 \\
111-11-111-1-11111 -11i1-111-1-1-1-11 \\
1111 -11-111-1-11111 -11i1-111-1-1-1-1 \\
11111 -11-111-1-11-111-11i1-111-1-1-1 \\
111111 -11-111-1-1-1-111-11i1-111-1-1 \\
-1111111 -11-111-1-1-1-111-11i1-111-1 \\
-1-1111111 -11-111-1-1-1-111-11i1-111 \\
1-1-1111111 -11-111-1-1-1-111-11i1-11 \\
11-1-1111111 -11-111-1-1-1-111-11i1-1 \\
-111-1-1111111 -11-111-1-1-1-111-11i1 \\
1-111-1-1111111 -11-111-1-1-1-111-11i
\end{pmatrix}$$

Fig 5: Matrix-5

$$\begin{bmatrix}
i1-1-11111-11-11-1-1-1-1-1-1-1111-11111-11111-11-1-1-1111111-1-11-11-11 \\
1i1-1-11111-1-1-11-1-11-1-1-1-1111-11111-11-111-111-1-111-1111-111-11-1 \\
-11i1-1-11111-1-1-11-1-11-1-1-11111-11111-11-111-111-1-11-1-1111-111-11 \\
-1-11i11-1111-1-1-1-11-1-11-1-1-11111-11111-11-111111-1-11-1-1111-111-1 \\
1-1-11i11-1111-1-1-1-1-1-1-11-11-11111-11111-11-11-1111-111-1-11-11-111 \\
11-111i1-1-11111-11-11-1-1-1-1-1-1-111-11-1111-11111-11-1-1-1111111-1-1 \\
111-111i1-1-11111-1-1-11-1-11-1-1-1-111-11-1111-11-111-111-1-111-1111-1 \\
1111-1-11i1-1-11111-1-1-11-1-11-1-1-1-111-111111-11-111-111-1-11-1-1111 \\
-11111-1-11i11-1111-1-1-1-11-1-11-1-11-111-1-11111-11-111111-1-11-1-111 \\
1-11111-1-11i11-1111-1-1-1-1-1-1-11-1-11-1111-11111-11-11-1111-111-1-11 \\
-1-1-1-1111-111i1-1-11111-11-11-1-1-1111-1-11-11-1111-11111-11-1-1-1111 \\
1-1-1-1-1111-111i1-1-11111-1-1-11-1-1-1111-111-11-1111-11-111-111-1-111 \\
-11-1-1-11111-1-11i1-1-11111-1-1-11-1-1-1111-111-111111-11-111-111-1-11 \\
-1-11-1-1-11111-1-11i11-1111-1-1-1-111-1-1111-111-1-11111-11-111111-1-1 \\
-1-1-11-11-11111-1-11i11-1111-1-1-1-111-1-11-11-1111-11111-11-11-1111-1 \\
-11-1-1-1-1-1-1-1111-111i1-1-11111-11-1-1111111-1-11-11-1111-11111-11-1 \\
-1-11-1-11-1-1-1-1111-111i1-1-11111-11-1-111-1111-111-11-1111-11-111-11 \\
-1-1-11-1-11-1-1-11111-1-11i1-1-1111111-1-11-1-1111-111-111111-11-111-1 \\
-1-1-1-11-1-11-1-1-11111-1-11i11-1111111-1-11-1-1111-111-1-11111-11-111 \\
1-1-1-1-1-1-1-11-11-11111-1-11i11-111-1111-111-1-11-11-1111-11111-11-11 \\
111-11-11-1-1-1-1-1-1-1111-111i1-1-1111-11-1-1-1111111-1-11-11-1111-111 \\
1111-1-1-11-1-11-1-1-1-1111-111i1-1-1-111-111-1-111-1111-111-11-1111-11 \\
-11111-1-1-11-1-11-1-1-11111-1-11i1-11-111-111-1-11-1-1111-111-111111-1 \\
1-1111-1-1-1-11-1-11-1-1-11111-1-11i1-11-111111-1-11-1-1111-111-1-11111 \\
11-1111-1-1-1-1-1-1-11-11-11111-1-11i1-11-11-1111-111-1-11-11-1111-1111
\end{bmatrix}$$

Performance Enhancement of TCP for Wireless Network

GJCST Classification
C.2.5, C.2.1

[1]Pranab Kumar Dhar, [2]Mohammad Ibrahim Khan,[3]P.M. Mahmudul Hassan

***Abstract*-Transmission Control Protocol (TCP) is one of the core protocols of the Internet Protocol Suite which provides reliable, ordered delivery of stream of bytes from a program on one computer to another program on another computer. TCP assumes congestion which is the primary cause of packet loss and uses congestion control mechanisms such as Tahoe, Reno and New Reno to overcome this congestion in wireless network. These TCP variants take longer time to detect and recover packet loss. In order to improve retransmission scheme, we propose a modified version of New Reno that outperforms previous TCP variants because of utilizing faster retransmission scheme as well as transferring more packets to the destination.**

Keywords-TCP/IP, Wireless Network, Transport protocol, Internet, Congestion Control.

I. Introduction

Today Internet is different from a single network. Because many parts in the world have different topologies, bandwidths, delays, packet sizes, and other parameters. TCP is a connection-oriented packet transfer protocol that ensures communication between two hosts. The idea behind this reliability and ordered packet delivery is that the sender does not send a packet unless it has received acknowledgment from the receiver that the previous packet or group of packets which already sent has been received [1-2]. Because of the good performance of TCP, most networks of current traffic use this transport service. It is used by the applications such as telnet, World Wide Web (www), ftp (file transfer protocol), and e-mail [3-4].Because of wide use of Internet as well as the widespread use of TCP by the majority of the network applications, it should be needed to improve the congestion detection and avoidance mechanism of TCP. Starting from the series congestion collapse on October 1986, many researchers developed and implemented different versions of TCP such as TCP Tahoe and TCP Reno and TCP New Reno [2]. In this paper, the mechanisms used by TCP Reno and TCP New Reno for controlling congestion on the network, are studied and analyzed using a Network Simulator known as NS2 [7-9]. From the simulation results we observe that TCP New Reno provides better performance than TCP Reno.Thus, in order to enhance the performance of TCP during network congestion, we propose a modified version of New Reno.The rest of the paper is organized as follows. Section II presents the background information regarding TCPvariants. Section III introduces our proposed modified New Reno. Section IV summarizes and discusses the experimental results and compares the performance of our Proposed New Reno with Reno and New Reno. Finally, section V concludes this paper.

II. TCP Variants

TCP is a component of TCP/IP Internet protocol suite. However, it is definitely considered as an independent, general purpose protocol since it can be used by other delivery systems.TCPprotocol makes very minor assumptions about the underlying network, for example it is possible to employ TCP over single network just like an Ethernet or even over complex networks such as the global Internet [5]. It is the dominant transport protocol over both wired and wireless links [6]. At the end of 1980s, a congestion control algorithm was proposed by Van Jacobson which is known as TCP Tahoe and today‘s Internet stability is mainly based on it. The first modified version of TCP Tahoe was the TCP Reno and then followed by other flavors or variants. The design of congestion control algorithm developed by Van Jacobson is based on the end-to-end principle and has been fairly successful from keeping the Internet away from congestion collapse. The following is a list that shows some of the TCP flavors or variants.

1) *Tahoe TCP*

The Tahoe TCP algorithm includes Slow-Start, Congestion Avoidance and Fast Retransmission. This algorithmincludes a modification to the round-trip time estimator usedto set retransmission timeout values. This is notsuitable for high band-width product links because it takes acomplete timeout interval to detect a packet loss and in fact, in most implementations it takes even longer time becauseof the coarse grain timeout.

2) *Reno TCP*

Reno improves the performance of Tahoe by introducing a Fast Recovery Phase. This phase activates after fast retransmission when three duplicate packets are lost. The parameters are initialized as follows:

Slow start threshold = 1/2 Congestion window;
Congestion window = Slow start threshold;
The Reno TCP works as follows:

(i) Slow start threshold and congestion window are both resized to reduce the transmission rate and drain the network congestion. In addition, we do not need to begin from slow start again.

Corresponding author[1]-Assistant Professor Department of Computer Science and Engineering Chittagong University of Engineering and Technology (CUET) Chittagong-4349, Bangladesh Cell: +88-01818-000112

About[2&3]- Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET), Chittagong-4349, Bangladesh

(ii) Until a non-duplicate packet' is received, a temporary congestion window is used during Fast Recoveryand isincreased by one message for every new received duplicatepacket; this allows new packet transmissions during Fast Recovery operation.

(iii) When the sender receives acknowledgment about the retransmission of the lost packets which are received successfully, the Fast Recoveryphase is terminated. Then CongestionAvoidancephase is initialized and congestion windowsize starts to grow from its updated congestion window size.But the problem with TCP Reno is that if the drop packets are multiple then the first information about the packet loss comes when receiver receives the duplicate acknowledgments. But the information about the second packet which is lost will come only when the sender receives the acknowledgment for the retransmitted first packet after one Round Trip Time (RTT). Thus, it does not provide good performance when multiple packets are dropped from a single window of data [6].

3) *New Reno TCP*

To overcome the limitations of TCP Reno, New Reno has been introduced. New Reno can be able to detectmultiple packet losses. For this reason, it is muchmore efficient than Reno in case ofmultiple packet losses.Like Reno, New Reno also entersinto fast retransmission process when it receives multiple duplicate acknowledgments. However, it differs from Reno when it does not recover fast until the acknowledgement is received. Thus, it overcomes the problem of Reno by reducing the congestion window in multiples times.The fast-transmit phase of New Reno is similar as Reno. The difference is that New Reno allows multiple re-transmissions in the fast recovery phase.When New Reno enters in fast recovery phase it calculates the maximum outstanding segment. The fast-recovery phase proceeds like Reno, however when a fresh acknowledgement is received, it considers two cases:

(i) If it sends acknowledgement to all outstanding packets, in that case it exits from Fast Recovery phase and sets congestion window to slow start threshold and continues to process congestion avoidance like Tahoe.

(ii) If the acknowledgement is partial in that case, it indicates that the next packet that is in line has lost and it re-transmits that packet again and sets the number of received duplicate acknowledgement to zero.

It exits from fast recovery stage when it sends acknowledgement to all the data available in the window.

The main problem of New Reno is that it takes one RTT to detect each packet loss. When the acknowledgement for the first retransmitted packet is received, after then we can detect the other lost packets.

III. Proposed Modified New Reno TCP

Proposed modified New Reno extends the retransmission mechanism of New Reno. It keeps track the packet transmission and it also estimates the RTT by calculating the time needed to get acknowledgment from the receiver. When a duplicate acknowledgment is received it calculates the time difference between current time and packet transmission time. If it is greater than RTT, then it immediately retransmit the packet without waiting for three duplicate acknowledgments or a coarse timeout. In order to recover multiple packets drop, the modified New Reno records the highest sequence number of the packet in a single window before retransmit the lost packet. If acknowledgment of the retransmitted packet does not cover the highest sequence number of the packet in a single window, then retransmit the indicated packet again. The remaining procedures of New Reno are unchanged in Modified New Reno.

IV. Simulation Results and Discussion

1) *Design of Simulation Structure*

In this section, we design the general process of the simulation including the topology of the Network and the simulation program to be used. Figure 1 shows the over-all simulation process.

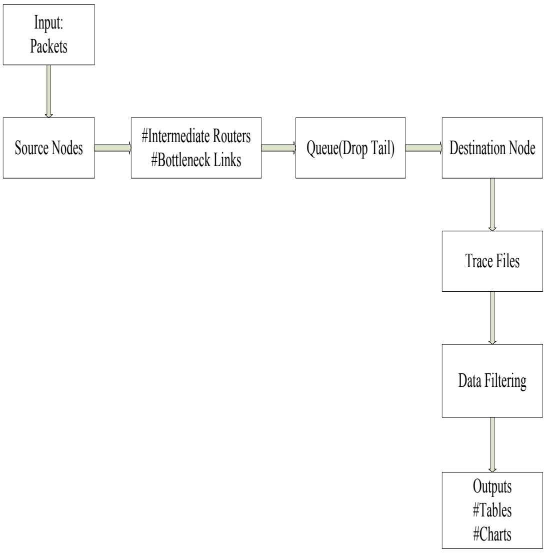


Fig. 1. Block diagram of the over-all process of the simulation

2) *Design of Network Topology*

The three algorithms of TCP congestion control mechanisms such as TCP Reno, TCP New Reno, and TCP Modified New Reno are represented and evaluated using the following network topology.

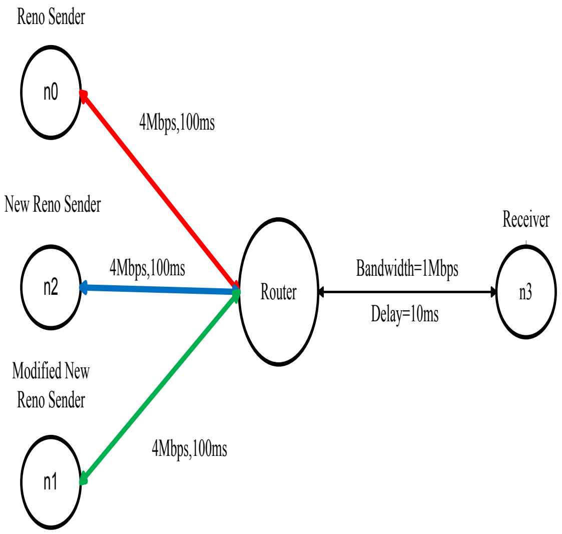


Fig. 2. Topology layout used for simulation

Using the above topology layout, three different sources of TCP are allowed to send FTP traffic to the destination. Each of the TCP source has a connection link of 4Mbps, (Mega bits per second), bandwidth and 10ms, (milli-seconds), delay to its nearest router, and Router. The bottleneck link from Router to node 3 (n3) has a bandwidth of 1Mbps and 10ms delay time. The bottleneck is suitable for evaluating the performance of the algorithms for congestion control and congestion avoidance mechanisms.

As all the sources should be pass through a Router, the queue size is bounded to a limit of 4(not fixed), or above and Drop-Tail queuing mechanism decides which packets will be discarded. Each TCP source is allowed to send FTP traffic with packet size of 460 bytes and the maximum queue size is 30 packets for the duration of 30 sec. Such traffic setup is suitable for collecting data from simulation on a chosen interval over a given bandwidth configuration.

In this paper, simulation results are measured and analyzed using some performance metrics such as number of received packets, number of acknowledgements, number of dropped packets, and throughput.In this simulation, the total packet size is 460 bytes and the maximum queue limit is 30 and also the duration of total simulation is 30 sec. Now if we increase the queue limit more than 30, then more packets will be waited in queue and delay of packet transmission will be increased. This is because we can not transmit all packets to the receiver with in 30 sec. In order to obtain better output we should keep queue limit as small as possible.From Figure 3, we observed that when queue limit is 4 our proposed Modified New Reno obtains the maximum number of received packets. If the queue limit is small, all TCP variants including our proposed New Reno also drop packets as shown in Figure 4. In our proposed New Reno, we have used a new retransmission mechanism that quickly recovers the lost packets, and almost all packets are transmitted to the receiver with in 30 sec. Thus, our proposed New Reno achieves higher number of received packets than Reno and New Reno. When queue limit is 8 the time delay of packet transmission is increased than previous queue limit.

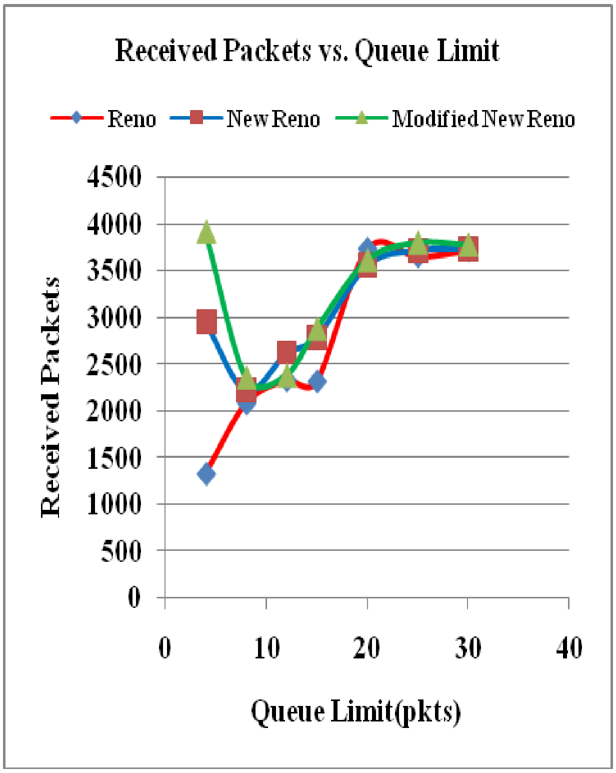


Fig. 3. Received packets vs. Queue limit

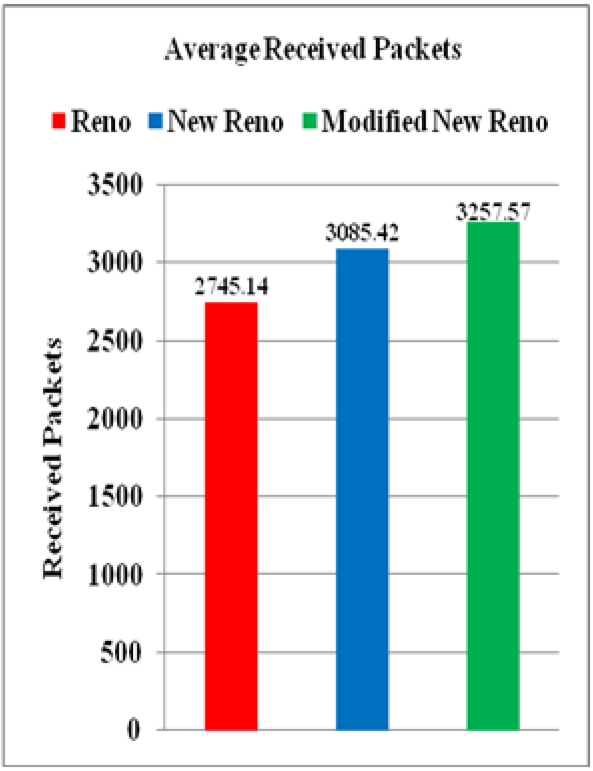


Fig. 4. Comparison of average received packets

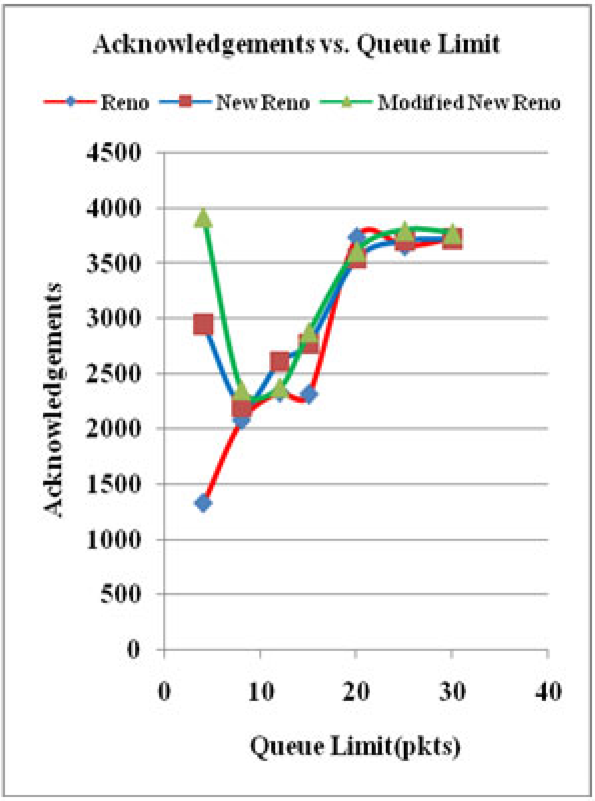


Fig. 5. Acknowledgements vs. Queue limit

Our proposed system has faster retransmission mechanism; however, it has higher delay time for packet retransmission which causes network congestion as like Reno and New Reno. For this reason all packets can not be transmitted to the destination within 30 sec. Thus we achieve lowest number of received packets. When queue limit is 12, New Reno achieves higher output because of its steady retransmission scheme. When queue limit is 15, our proposed New Reno achieves higher number of received packets. When queue limit is 20 as compared to total packet size of 460 bytes, the delay of packet transmission in queue is increased and the three TCP variants have almost same number of received packets. For queue limit of 25, the three TCP variants have all most same output but our proposed New Reno achieves a little bit higher output than previous TCP variants and these outputs will be continued for higher queue limits.

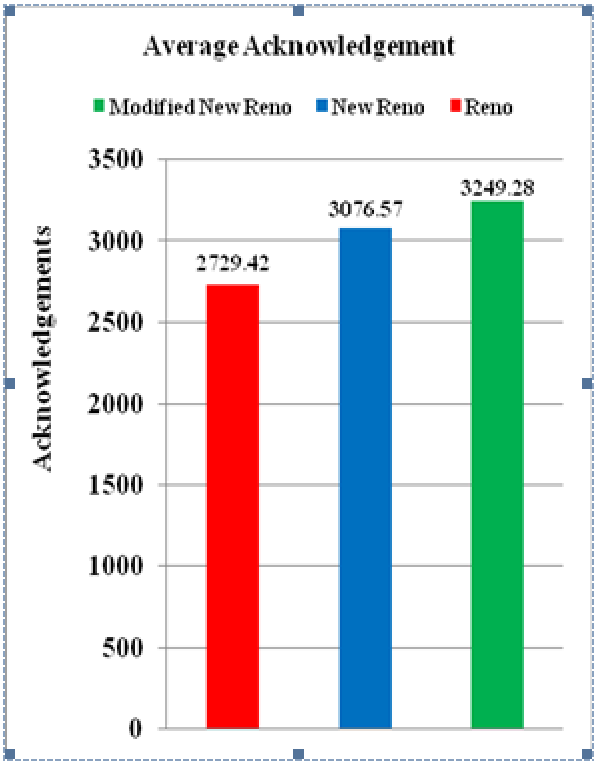


Fig. 6. Comparison of average acknowledgement

Thus, we can say that if we use large queue limit then the delay of packet transmission from queue is increased. For this reason, we cannot transmit all packets to receiver with in 30 sec as compared to queue limit of 4 in which we can transmit maximum number of packets. Thus, we should limit queue size as small as possible to obtain better output.

From Figure 5 it is seen that acknowledgements send by receiver at different queue limits using Modified New Reno is greater than Reno and New Reno. Figure 6 shows that average acknowledgements send by receiver using Modified New Reno is higher than Reno and New Reno.Figures 7 and 8 show that the number of dropped packets using Modified New Reno is less than Reno and New Reno. If we consider the throughput (Good-put) of Reno, New Reno and Modified New Reno as shown in Figure 9, we observe that the throughput at different queue limits using Modified New Reno is better than Reno and New Reno. If we calculate the average throughput as shown in Figure 10, we also observe that Modified New Reno has better average throughput than Reno and New Reno.Thus, we can conclude that our proposed Modified TCP New Reno speeds up the performance of TCP by improving the end-to-end throughput.

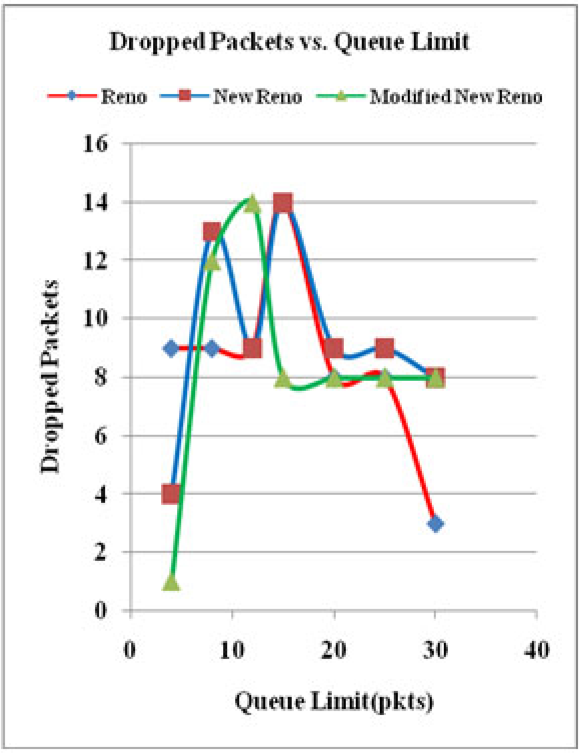


Fig. 7. Queue limit vs. Dropped packets

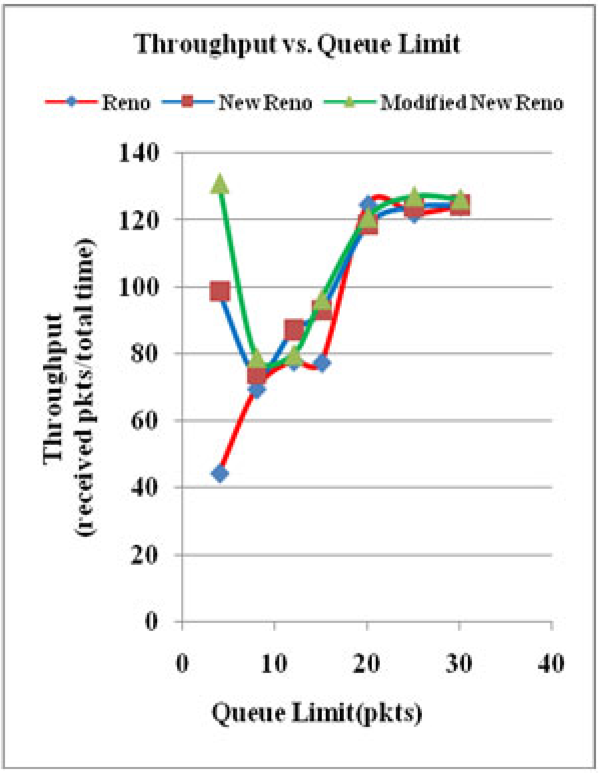


Fig. 9. Throughput vs. Queue limit

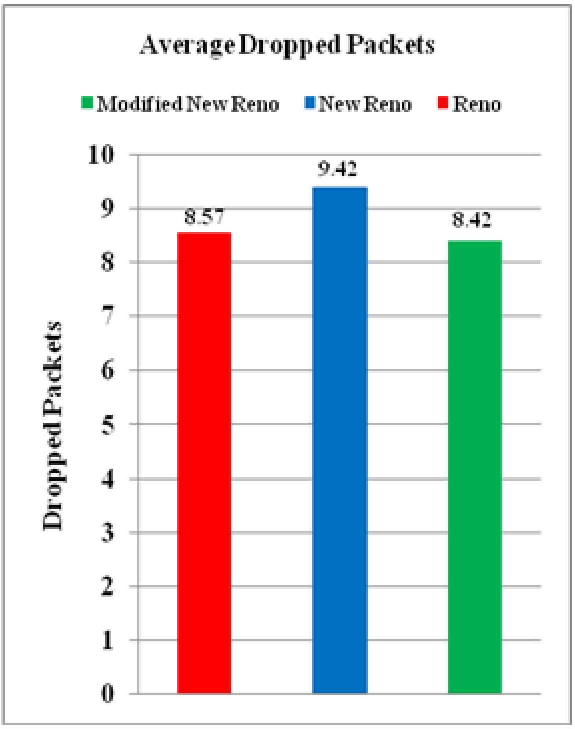


Fig. 8. Comparison of average dropped packets

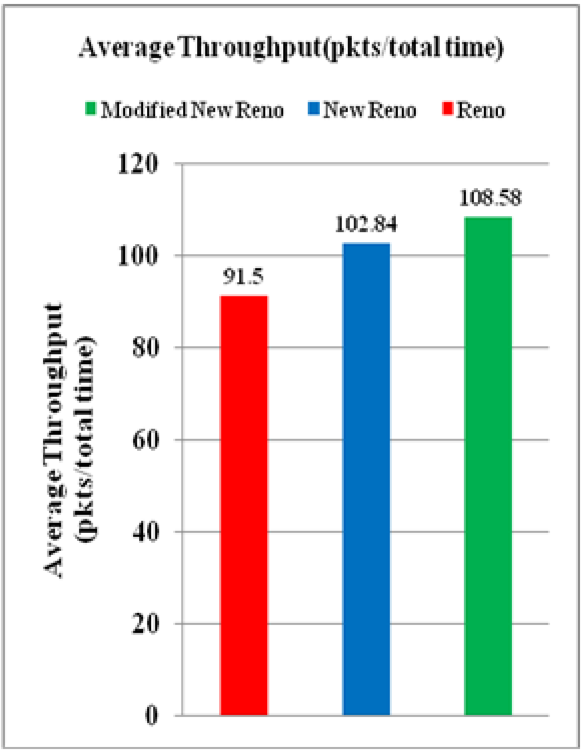


Fig. 10. Comparison of average throughput

V. Conclusion

In this paper, we have presented a modified version of New Reno to improve the TCP performance in wireless network when congestion occurs. Simulation results indicate that Modified New Reno outperforms Reno and New Reno in terms of received packets, dropped packets and throughput. This is because our proposed scheme does not have to always wait for 3 duplicate acknowledgements. For this reason, it can retransmit quickly and does not reduce the window size too much as like Reno. In addition,it prevents many of the coarse grained timeouts of New Reno as it does not need to wait for 3 duplicate acknowledgments before it retransmits a lost packet. Finally, we can say that our proposed New Reno can be a suitable candidate for TCP congestion control.

VI. References

1) Chuck Semeria, "Supporting Differentiated Service Classes: TCP Congestion Control Mechanisms," Juniper Networks, 2000.
2) Van Jacobson and Michael J. Karels, "Congestion Avoidance and Control," November, 1988.
3) Transmission Control Protocol (TCP) http://www.erg.abdn.ac.uk/users/gorry/course/inet-pages/tcp.html
4) TransmissionControlProtocol.http://en.wikipedia.org/wiki/Transmission Control_Protocol.
5) D.E.Commer, Internetworking TCP/IP: Principles, Protocols, and Architecture, Upper saddle, New Jersey: Prentice Hall, 2006.
6) M. Ghaderi TCP Aware Resource Allocation in CDMA Networks" In Proc. ACM MobiCom, Los Angeles, USA. 2006.
7) Network Simulator (NS-2) web site: http://www-mash. cs.berkeley.edu/ns
8) The Network Simulator (NS-2) January 6, 2009.
9) Thenetworksimulator-ns-2, http://www.isi.edu/nsnam/ns

An Efficient Technique for Iris Data Compression an Algorithm By Bezier Curve Approach

Piyush Tripathi[1] , Aparna Shukla[2]

GJCST Classification
E.4

***Abstract*-Iris is a strong biometric tool used for human authentication. In this methodology, the patterns in the iris such as rings, furrows and freckles can be envisaged a set of Bezier curves and hence represented by the corresponding Bezier points, resulting in considerable reduction in the file size. After the iris is captured using scanner, the patterns are extracted from the scanned image. Then they are treated as a Bezier curve and the coordinated of the characteristics four control points are determined. These set of coordinates of the control points are stored as a data file representing the iris, resulting in a considerable memory saving. Whenever the iris, is needed for recognition the retinal blood vessels can be regenerated by drawing the corresponding Bezier curves using the control points. Truthfulness of this regenerated iris is ascertained mathematically.**

Keywords-Iris; Mapping and Regeneration; Bezier curves; compression, cross correlation coefficient

I. Introduction

Iris' are composed before birth and, except in the event of an injury to the eyeball, remain unchanged throughout an individual's lifetime. It is a membrane in the eye. Iris is a biometric attribute which can be used for authenticating and distinguishing people. The pattern extracted from the iris is unique for even genetically identical twins.Iris patterns are extremely complex; carry an astonishing amount of information. The iris-scan process begins with a photograph. A specialized camera, typically very close to the subject, no more than three feet, uses an infrared imager to illuminate the eye and capture a very high-resolution photograph. This process takes only one to two seconds and provides the details of the iris that are mapped, recorded and stored for future matching or verification. The iris can be combined with any authentication factor and can be used as a powerful tool against repudiation.

1) *Data Reduction*

Iris recognition technology converts the visible characteristics as a phase sequence into an Iris code. Usually the size of the template is 512 bytes. A template stored, is used for future identification attempts. Here in this work, a methodology is presented, by which an iris can be stored with in a memory space of about 100 to 200 bytes only, resulting in considerable reduction in data size and from which an acceptable quality of an iris can be regenerated. The details corresponding to the rings, furrows and freckles. are carefully retained to the maximum extend, so that the loss of information is minimized

2) *Bezier Representation*

In an attempt to achieve data reduction in storing the patterns of an iris, each rings, furrows and freckles are treated as a Bezier curve. A Bezier curve is a parametric curve important in computer graphics. Bezier curves were widely used to designed automobile bodies. The curves can conventionally be represented by de Casteljau's algorithm. A Bezier curve is a function of four control points, of which two will be the two end points lying outside the curve. These four points completely specify the entire curve [11]. The curve can be regenerated uniquely, from the control points. Each and every pattern of the iris being treated as a Bezier curve and by using the Bezier equation, the end points and the control points are determined. Thus every pattern in the iris gives raise to four Bezier points. Thus all are represented as Bezier points. So that instead of storing the entire iris in template, only the collections of Bezier points are stored. Whenever the iris is needed, it can be regenerated as a set of Bezier curves using this set of control points.

II. The Algorithm

The Bezier equation of a curve being

$$P\ \ u\ \ = \sum_{k=0}^{n} P_k J_{k,n}\ \ u\ \ , 0 \le u \le 1$$

Where $P_k = (X_k, Y_k, Z_k)$, K=0 to n are used to produce the position vector P(u) on the path of an approximation Bezier polynomial function between P(0) and P(n).
The x co-ordinate of any point on a Bezier curve is given by

$$x\ \ u\ \ = \sum_{k=0}^{n} X_k J_{k,n}\ \ u\ \ , 0 \le u \le 1$$

$$\text{where } P_k = X_k$$

and similarly the y coordinate of any point is represented by

$$y\ \ u\ \ = \sum_{k=0}^{n} Y_k J_{k,n}\ \ u\ \ , 0 \le u \le 1$$

$$\text{where } P_k = Y_k$$

About[1]- Amity School of Engineering and TechnologyViraj Khand-5, Lucknow, UP(India) Email:-piyush.tripathi2007@gmail.com
About[2]- Birla Institute of Technology B-7, Industrial Area Post-TSL, Naini-Allahabad,UP(India)Email:-meet2aparna@gmail.com

But the same x(u) and y(u) can also be obtained, as given below, from a unique set of the four control points (x_0, y_0), (x_1, y_1),(x_2, y_2) and (x_3, y_3) where the first and the last points are the two end points of the given Bezier curve.

$$x = x\ \ u\ \ = x_0\ \ 1-u\ ^3 + 3x_1.u\ \ 1-u\ ^2$$
$$3x_2.u^2\ \ 1-u\ \ + x_3.u^3\ .....(1)$$

$$y = y\ \ u\ \ = y_0\ \ 1-u\ ^3 + 3y_1.u\ \ 1-u\ ^2$$
$$3y_2.u^2\ \ 1-u\ \ + y_3.u^3\ .....(2)$$

The present work treats each pattern of an iris as a Bezier curve and four control points are extracted. Hence, it is sufficient to store these four control points instead of storing thewhole pattern. At the user end the patterns can be reproduced from these control points.

1) *Obtaining the control points*

In the iris, each and every pattern further is considered individually as a Bezier curve to find the control points. It is divided in to n equal intervals. Then the deviation of $x_0, \Delta x_0$ and $x_3, \Delta x_3$ and y_0, Δy_0 and $y_3, \Delta y_3$ Is taken from these values, the slope of the tangent at the coordinate (x_0, y_0) and (x_3, y_3) is manipulated. Using the slope and y_0 and y_3, the straight-line equations of the tangent in both the endpoints are fitted as shown in figure 1. Then the control points x_2 and x_3 are initialized to x_0 and x_3 and varied with the step value depending on the number of divisions of the curve. These are the assumed control points for x_1 and x_2. These values are substituted in the tangent line made at the end points and the assumed y values are computed. Because the second and third control points are located at the tangent lines made at the end points of the curve [1,2]. Now from the above equation (1) and (2), the u values are substituted 0.2 and 0.8; the following conjucate equations are obtained

$$x = x\ \ 0.2\ \ = 0.512x_0 + 0.384x_1$$
$$+0.096x_2 + 0.008x_3\ .....(3)$$

$$y = y\ \ 0.2\ \ = 0.512y_0 + 0.384y_1$$
$$+0.096\text{y}_2 + 0.008y_3\ .....(4)$$

$$x = x\ \ 0.8\ \ = 0.008x_0 + 0.096x_1$$
$$+0.384x_2 + 0.512x_3\ .....(5)$$

$$y = y\ \ 0.8\ \ = 0.008y_0 + 0.096y_1$$
$$+0.384\text{y}_2 + 0.512y_3\ .....(6)$$

In the above equations (3), (4), (5) and (6), the assumed control points along with x_0, x_3 and y_0, y_3 are substituted. Thus the assumed value at $x_{(0.2)}$, $y_{(0.2)}$, $x_{(0.8)}$ and $y_{(0.8)}$ are manipulated and checked with the original curve coordinates. If there is a match for both the points, the assumed control points are fixed as the actual control points of the curve. The above procedure can be summarized as two lemmas as below:

Lemma 1: In a Bezier curve with the starting and ending points at P_0 and P_3, the control

Lemma 2: For a certain combination the positions of a point A moving along the tangent at the starting point and a point B moving along the tangent at the end of the given Bezier curve, the presently generated curve will exactly fit in with the given Bezier curve, provided A and B approach each other.

Proof: Let for a given combination of the positions of A and B.

Let x‘ = x at u = m (0<m<1) and y‘ = y at u = m be computed. Likewise x‘‘ = x at u = 1-m and y‘‘ = y at u = 1-m be computed. Now since A and B approach each other during every iteration for every combination of their position, as per Lemma 1, there must be a Bezier curve passing through x‘ and x‘‘. Let the sum of the square error defined as. SSE and y‘a and y‘‘a are actual y coordinates of the given curve. Then Δ

SSE = (y‘a - y‘)2 + (y‘‘a - y‘‘)2 ≥ξ.

where ξ is the maximum error radius that is permissible during numerical evaluation around the neighbourhood of y‘a and y‘‘a and when ΔSSE = ξ the presently generated Bezier curve exactly matches with the given curve for all values of m. Hence the algorithm for numerical evaluation of the control points P_1 and P_2, for a set of points forming any curve on the x-y plane can be as below:

Step 1. Put the x-y coordinates of the points on a furrow in an array

Step 2. Evaluate the equation-1 of the tangent to the curve passing through (x_0, y_0)

Step 3. Evaluate the equation-2 of the tangent to the curve passing through (x_3,y_3)

Step 4. Set the assumed x-coordinate xc_1 of the first control point at x_0

Step 5. Use equation-1, and compute the y-coordinate of the assumed first control point

Step 6. Set the assumed x-coordinate xc_2 of the second control point at x_3

Step 7. Use equation-2, and compute the y-coordinate of the assumed second control point

Step 8. Compute x, y values corresponding to u = 0.2 and 0.8 using Bezier‘s conjucate equations

Step 9. If corresponding to the actual x value of the curve equal to the computed value of the x at u=0.2, the actual y value agrees to the computed y-value with an error radius of e, then step 10 otherwise step 13

Step 10. If corresponding to the actual x value of the curve equal to the computed value of the x at u=0.8, the actual y value agrees to the computed y-value with an error radius of e, then step 11 otherwise step 13

Step 11. Assign (xc_1,yc_1) and (xc_2,yc_2) as the two original control points
Step 12. End
Step 13. Decrement xc_2 to the next value
Step 14. Go to Step 7
Step 15. Increment xc_1 to the next value
Step 16. Go to Step 5

This sequence of evaluations extract the two desired control points numerically from a set of x-y values on a curve, with equally spaced x-values, instead of equally spaced u-values.

2) *Multi-y-valued Furrows*

A portion of the patterns may have multiple values for y for the same value of x. For such a many valued curves, a ninety-degree rotation with respect to the coordinate system will make them single valued. Here just the x co-ordinates are changed into y co-ordinates and the y co-ordinates are changed into x co-ordinates provided they are stored after taking in to account this fact of rotation again implemented while storing.

3) *Multi segmented Furrows*

In the case of self-folding or non-trivial curves, it will be necessary to break the furrows in to two or more simple curves, each one of which can be represented as a Bezier curve. In general, a furrow can be visualized as being composed of with many segments depending up on its complexity. The algorithm treats each segment as a Bezier curve.

4) *Combination of multi-segmented and multi-y-valued*

A furrow having many multi-y-valued portions and also the self-coiling necessitates the segmentation. The furrows and freckles should be divided in to at least three segments, each segment becoming a well-behaved Bezier curve. A self-coiling or a closed or near-closed ridge is first split into multiple segments and each segment is treated as a Bezier curve. In case any of these segments are multi-y-valued then it is given a ninety-degree rotation before extracting the control points.0

III. Storing the Data Points

For a typical iris pattern, as shown in Figure.2(a), there are about 20 Bezier curves, each of which can be represented by 4 control points that is by means of 8 coordinates, requiring 8 bytes of memory space since it is stored as a BCD. Thus the entire information content of the iris can be stored with about 20 X 8 = 160 bytes. When want to regenerate the fingerprint by any user or application, it can be regenerated precisely using these control points alone. Thus this near non-lossy method is able to store the iris information in about 200 bytes.

IV. Regeneration of the Ridges from the Control Points

The stored Bezier co-ordinates were read from the file and then it is substituted in the Bezier equation (1), (2) and the x, y co-ordinates of the curve to plot every furrows and freckles of the iris. Samples of two irises are shown below. Figure.2(a) and 3(a) represents the original irises, 2(b) and 3(b) represents the extracted furrows, freckles and rings of the same. Then the regenerated irises are shown in Figures 2(c) and 3(c). Thus the patterns are extracted from the original iris shown in Figure 2(a). Then the patterns of the original iris as shown in Figure.2(b) is thus construed as Bezier curves as per the above discussed methodology, their corresponding control points were computed and the collection of these points are stored as the iris file. Then the regeneration of the irises are done using this file having the control points and the regenerated iris is shown in Figure 2(c).

V. Correlation of the Regenerated Iris Withoriginals

In order to evaluate to what extent these regenerated patterns truthfully represent original iris, cross correlation coefficient is evaluated. If an iris recognition algorithm is developed then to find the accuracy of the algorithm, Equal Error Rates of ID accuracy can be used. But this algorithm is for reduction of memory storage in iris and thus the cross-correlation coefficient is used. The cross-correlation coefficient is presented in Table.I.

These values suggest that the regenerated patterns agree very well with the original ones. The correlation is strongly positive as seen in the Table I. All the irises are chosen to have 130x120 pixels for the sake of uniformity. Then both the original and the regenerated files were read using Matlab6.5 and cross-correlated for verification. Depending upon the value of the cross correlation coefficient, the acceptability of the reproduced one is decided. As a general hypothesis, taking more number of _u' intervals yield better correlation. A complete examination should always be carried out to identify multi y-valued or multi segmented furrows and their combination, so that the deviations in the values of the cross-correlation coefficient are minimized. The Figure.4 depicts the graph, which represents the relationship between two original irises with the reconstructed irises.

VI. Related Work

The template methodology is used for representing irises. Early ideas are by Flom and Safir in the year 1987[10]. From that numerous techniques have been developed. Among that the most advanced system has been developed by Daugman[9]. His work involves a technique based on wavelet like technique. Similar technologies have been developed by performing pattern matching via Gabor filter banks[7] by combining the Hough transform for iris localization[6] and by matching multiscale iris representations[8]. Daniel and Kirovski introduced EyeCerts. It converts the iris using a modified Fourier-

Mellin transform into a standard domain where the common radial patterns are conscisely.

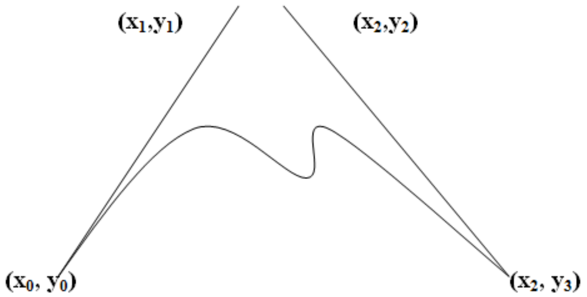


Fig 1. Sample curve along with the tangents made at the end points

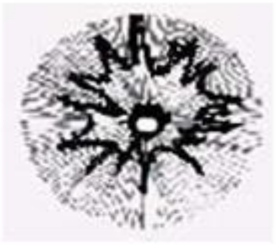


Fig.2(a).Original Iris

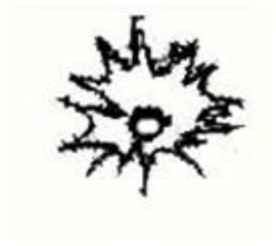


Fig.2(b). Extracted Furrows

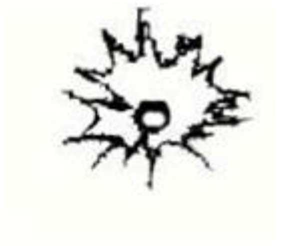


Fig.2(c). Reproduced Iris

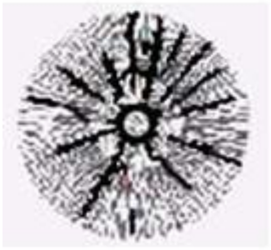


Fig.3(a). Original Iris

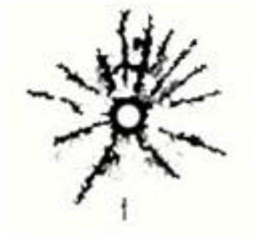


Fig.3(b). Extracted Furrows

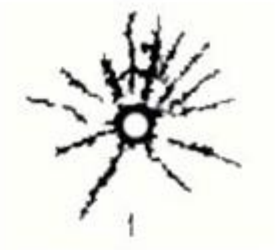


Fig.3(c). Reproduced Iris

Table I. The normalized cross correlation coefficients between the original and the reconstructed fingerprints

	Regenerated fingerprint from the Bezier control points of	
	Fingerprint 1 (R1)	Fingerprint 2 (R2)
Original 1	0.9961	0.3401
Original 2	0.2470	0.9999

represented[5].Basit represents eigen-irises after determining the centre of each iris and recognition is based on Euclidean distances[12]. Iris image is convolved with a blurring function which is a 2D Gaussian operator[11]. All the above algorithms require considerable memory space for storing an iris file. Similarly Bezier Curve is used in fingerprint technology for feature extraction and thus fingerprints matching [3]. For water marking the curves, the Bezier Blending function is used [4].

VII. Review of the Result

A C++ package is developed to generate Bezier control points using the (x,y) inputs of the iris furrows. In order to read the co-ordinates of each every furrow in the iris, XY-it software is used. The collections of all the Bezier control points are stored as a file. The file size is noted. A large reduction in the file size is observed. For the above-mentioned iris1, if it is stored in a JPEG format, it occupies 24.7KB. The extracted JPEG file occupies 18.5KB. When if it is stored as Bezier points, after dividing in to the necessary furrows, totally it has only 22 curves to be stored. The entire file size is exactly 176 bytes. Similarly fro the second iris, the original file size is 25.4KB in JPEG format. The extracted JPEG file occupies 18.5KB. But in Bezier representation, totally it has only 19 curves and it occupies exactly 152 bytes. The iris generation from the control points program is also written in C++. The graphics output of this particular program can be converted in to JPEG file if needed. In order to check for its truthfulness, the cross correlation coefficient is estimated. For that a program was developed using Mat Lab6.5.

VIII. Conclusion

In this paper, a methodology is proposed to store irises as a collection of Bezier control points resulting in the finite saving of the storage space. This also results in reduced time overhead to store, retrieve and compare irises.

IX. References

1) Md. Al-Amin Bhuiyan and Hiromitsu Hama, An Accurate Method for Finding the Control Points of Bezier Curves, Proceedings of the Osaka City University, 1997, Vol. 38, pp. 175-181
2) Md. Al-Amin Bhuiyan and Hiromitsu Hama, Identification of actors drawn in Ukiyoe Pictures, Pattern Recognition, 2002, Vol. 35, pp. 93102
3) Yuvan Huaqiang, Ye Yangdong, Deng Jianguang, Chai Xiaoguang and Li Yong, A Fingerprint feature extraction algorithm based on curvature of Bezier curve, Progress in Natural Science, 2007, Vol. 17, pp. 13761381
4) Yongjian Hu, Heung-Kyu Lee and Huafei Zeng, Curve Watermarking Technique for Fingerprinting Digital Maps, Proceedings of the IEEE 2008 International Conference on ntelligent Information Hiding and Multimedia Signal Processing, 2008, Issue 15-17, pp. 223-226
5) Daniel Schonberg and Darko Kirovski, EyeCerts, IEEE Transactions on Information Forensics and Security, June 2006, Vol.1, No.2,pp. 144-153
6) C.I. Tisse et al., Person Identification Technique using Human Iris Recognition",J.Syst. Res., 2003, Vol. 4, pp. 67-75
7) L. Ma, Y.Wang and T.Tan, Iris Recognition based on Multi Channel Gabor Filtering", Proc. Asian Conference of Computer Vision, 2002, pp.23-25
8) R.P.Wildes, Automated Iris Recognition : An Emerging Biometric Technology", Proc. IEEE, Sep.1997, Vol.85, No.9, pp. 1348-1363
9) J.Daugman, Recognizing persons by their Iris Patterns", Biometrics : Personal Identification in Networked Society, 1999
10) L.Flom and A.Safir, Iris Recognition System", U.S. Patent 4641 349, Feb.3, 1987
11) Y.P.Huang, S.W.Luo and E.Y.Chen, An efficient iris recognition system", Proc.1st Int. Conf. Mach. Learning and Cybermetics, Beijing, Nov 2002, pp.4-5
12) A.Basit, M. Y. Javed and M. A. Anjum, Efficient Iris Recognition Method for Human Identification", World Academy of Science, Engineering and Technology, 2005, pp. 24-26
13) L.Ma, T.Tan, Y.Wang, D.Zhang, Efficient iris recognition by characterizing key local variations" IEEE transactions on Image processing, June 2004, vol. 13, no. 6
14) Steven Harrington, Computer Graphics-A Programming Approach, McGraw – Hill International Editions, New York (1987)
15) Amos Gilat. Mat Lab An Introduction with Applications, Wiley India Pvt. Ltd, New Delhi (2007)
16) N. Krishnamurthy, Introduction to Computer Graphics, Tata McGraw – Hill Publishning Company Ltd., New Delhi (2006)

Applications and Scope of Virtualization for Server Consolidation in IT Industry

Gurjit Singh Bhathal [1], Dr. G.N. Singh[2]

GJCST Classification
B.m, C.m

***Abstract*- Virtualization is a broad term that refers to the abstraction of computer resources. In simple and useful definition, "Virtualization is a technique for hiding the physical characteristics of computing resources from the way in which other systems, applications, or end users interact with those resources. This includes making a single physical resource (such as a server, an operating system or storage device) appear to function as multiple logical resources, or it can include making multiple physical resources appear as a single logical resource."Platform virtualization: Platform virtualization is performed on a given hardware platform of host computer with virtual machine, which creates a simulated computer environment for its "guest" software.Resource virtualization:Resource virtualization is the virtualization of specific system resources, such as storage volumes, name spaces, and network resources.This paper will discuss their types, applications and advantages of virtualization technologies. It will also demonstrate the business problems.**

Keywords- Virtualization, Virtual Machine, VMware.

I. Introduction

Virtualization has a long history, starting in the mainframe environment and arising from the need to provide isolation between users. The basic trend started with time-sharing systems (enabling multiple users to share a single expensive computer system), aided by innovations in operating system design to support the idea of processes that belong to a single user. The addition of user and supervisor modes on most commercially relevant processors meant that the operating system code could be protected from user programs, using a set of so-called "privileged" instructions reserved for the operating system software running in supervisor mode. Memory protection and, ultimately, virtual memory were invented so that separate address spaces could be assigned to different processes to share the system's physical memory and ensure that its use by different applications was mutually segregated. These initial enhancements could all be accommodated within the operating system, until the day arrived when different users, or different applications on the same physical machine, wanted to run different operating systems A number of important challenges are associated with the deployment and configuration of contemporary computing infrastructure. Given the variety of operating systems and their many versions - including the often-specific configurations required accommodating the wide range of popular applications - it has become quite difficult to establish and manage such systems. This requirement could be satisfied only by supporting multiple VMs, each capable of running its own operating system.Significantly motivated by these challenges, but also owing to several other important opportunities it offers, virtualization has recently become again a principal focus for computer systems software. It enables a single computer to host multiple different operating system stacks, and it decreases server count and reduces overall system complexity. VMware is the most visible and early entrant in this space, but more recently XenSource, Parallels, and Microsoft have introduced virtualization solutions. Many of the major systems vendors, such as IBM, Sun, and Microsoft, have efforts under way to exploit virtualization. Virtualization appears to be far more than just another ephemeral marketplace trend. It is poised to deliver profound changes to the way that both enterprises and consumers use computer systems.

II. Business problems

The IT industry has dramatically evolved over the last decade, allowing businesses to gain access to technology through inexpensive x86 server systems, as well as the applications and operating systems that run on this platform. However, the adoption rate has grown so rapidly that many customers today are forced to deal with the following issues:

1) *Inefficient server Hardware Migration*

The typical enterprise replaces servers every three years. Although replacing servers may seem like a straightforward process, it can be quite time-consuming, painful, and expensive. The main issue surrounding this is the fact that each operating system is tied directly to the hardware, thereby making it difficult to migrate to newer servers, also in some instances applications can be tied to a particular named instance of the operating system, as well as the hardware. Therefore, each infrastructure refresh cycle can be unattractive to the datacenter, operating systems and applications management teams.

2) *Inefficient Application Server Deployment*

Application servers are continually added to enhance the business; however, lead times for procuring the hardware and software, performing testing and development, and conducting proof-of-concept modeling, implementation, and end-user training can sometimes take months to complete. Although a number of server deployment technologies are available, they can be very expensive to purchase and

About[1]- Department of Computer Science, Principal, Bhai Gurdas Polytechnic College, Sangrur-148001 (P.B) India (Telephone: +91-9814205475 email: gurjit.bhathal@gmail.com)

About[2]- HOD, Department of Physics and Computer Science, Sudarshan Degree College Lalgaon, distt. Rewa (M.P) India (Telephone:01755181208 email: gnsingh_27@rediffmail.com)

implement, which is why they are not more prevalent in businesses today.

3) High Availability Complexity

Due to the various application architectures and operating systems available, high availability can be difficult to implement. In general, we look at high availability as a series of measures undertaken to implement minimal to near real-time failover for a particular application *within* the data center. Implementing a highly-available infrastructure increases in complexity as the size of the data center grows, which is what makes it expensive to implement and maintain. Most customers do not need every system to be highly-available, but in general, systems that serve the network backbone, directory services, file and print sharing, email, enterprise applications etc. generally fall into the _high availability' category. Determining the criticality of each application is the first step in creating a highly available infrastructure. This determination should be made by upper management (not IT) and incorporated into the enterprise's Disaster Recovery Plan. Unfortunately, in almost every case it takes a major application outage to demonstrate the importance of high availability which can be avoided through regular planning and testing.

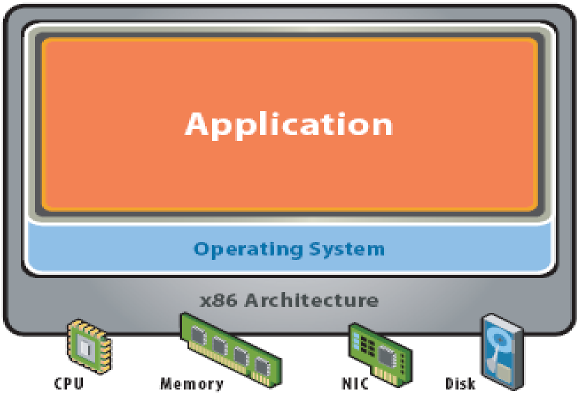
4) Disaster Recovery Complexity

Disaster Recovery planning has been a major focus for customers in light of recent terrorism activity and power grid failures, as well as the various natural disasters involving tropical storms, tornadoes, and hurricanes over the last decade. However, there are very few enterprises that have implemented disaster recovery procedures as well as a regular testing program. We look at Disaster Recovery as a series of measures undertaken to implement minimal to near real-time failover for a particular application *outside* the data center involving a hot or cold site. Similar to a highly-available infrastructure, creating a Disaster Recovery site increases in complexity as the size of the data center grows, which makes it incredibly expensive to implement and maintain. Businesses sometimes maintain multiple sites for Disaster Recovery, and in some cases, duplicate the entire infrastructure to avoid any recovery difficulties during large-scale recoveries. Some implement multiple storage area networks and replicate between sites asynchronously or synchronously, which can be very expensive and problematic due to network latencies and distance limitations. Other technologies used in these types of scenarios are geographically dispersed clusters, with nodes in multiple data-centers, giving customers the ability to fail over applications to different data centers at the push of a button. This is a very challenging technology to implement correctly, and in many cases is very expensive and difficult to maintain.

III. Different Applications and Models

1) Actual Physical Computing Model

Implementing a physical server with an operating system for each application is the most universally deployed server strategy. This strategy isolates the application and prevents any other applications from consuming resources, thereby enhancing application stability and ensuring a consistent end-user experience. However, this physical model actually accounts for the problems associated with server utilization and proliferation.

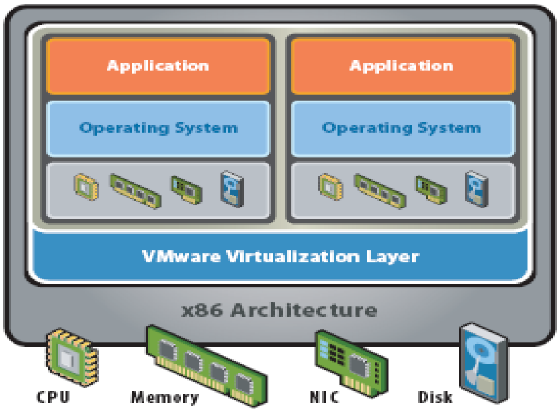


Before Virtualization:

Fig.1. Normal used Architecture of System.

2) Virtual Computing Model

In contrast to the physical model, the Virtual Computing Model increases server utilization while consolidating multiple workloads or server instances on a single physical system.



After Virtualization:

Fig.2. Virtual Architecture using software layer.

3) Server Virtual Computing Model

Over the last few years, a number of software and hardware vendors have entered the server virtualization space with a common mission of creating hardware independence, increasing utilization, and developing solutions to ease the migration to a real-time enterprise. In order to properly understand virtualization, let's take a look at two of the most common computing models today. Virtual machines are used to consolidate many physical servers into fewer servers, which in turn host virtual machines. Each physical server is reflected as a virtual machine "guest" residing on a virtual machine host system. This is also known as Physical-to-Virtual or 'P2V' transformation

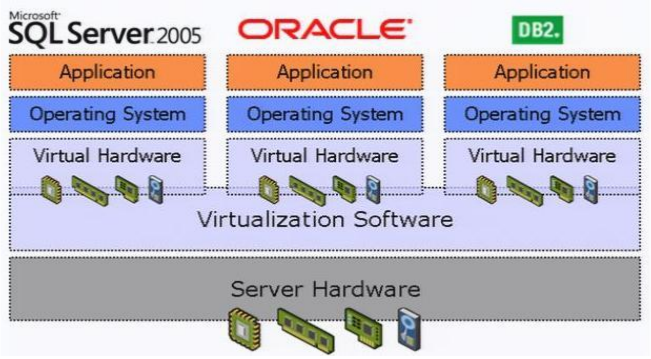


Fig.3. Server Consolidation on one Physical System.

4) Disaster recovery Simplicity

A virtual infrastructure eases the disaster recovery process. Since each virtual machine is independent from the physical hardware, and is encapsulated to a single file, virtual machines can be copied via SAN replication to another data center and can run on completely different hardware. This generally requires a high bandwidth connection between sites, but an alternative approach can also involve the traditional method of tape restoration. Figure 11 below illustrates how SAN replication can copy a virtualized environment from one site to another, thereby cloning the entire production environment in real time. Virtual machines can be used as "hot standby" environments for physical production servers. This changes the classical "backup-and-restore" philosophy, by providing backup images that can "boot" into live virtual machines, capable of taking over workload for a production server experiencing an outage.

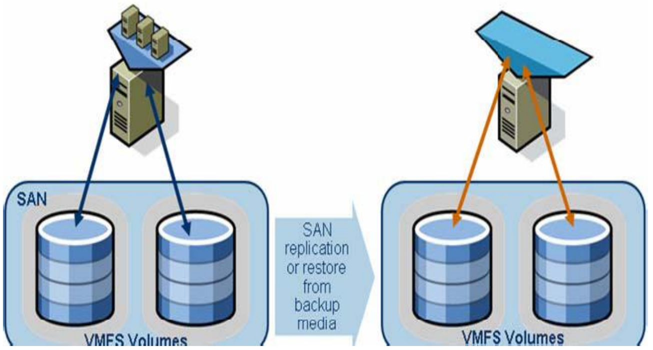


Fig.4.Disaster Recovery Architecture in a Virtual Environment by VMware

IV. Conclusion

The use of virtualization portends many further opportunities for security and manageability on the client. The examples presented here only begin to illustrate the ways in which virtualization can be applied. Virtualization represents a basic change in the architecture of both systems software and the data center. It offers some important opportunities for cost savings and efficiency in computing infrastructure, and for centralized administration and management of that infrastructure for both servers and clients. We expect it to change the development, testing, and delivery of software fundamentally, with some immediate application in the commercial and enterprise context.

V. References

1) Mendel Rosenblum, Chief Scientist, VMware Inc
2) Dr. Dobb‘s Journal August 2000 By Jason Nieh and Ozgur Can Leonard .
3) From Wikipedia, the free encyclopedia

4) Simon Crosby, XenSource and David Brown, Sun Microsyst

Three Dimensional Database: Creating Dynamic Web Controls Using XML and XSLT

R. Vadivel[1],Dr. K. Baskaran[2]

GJCST Classification
H.2.1, H.3.5

Abstract-**a dynamic web application is vital for online business. It has increased the on-demand needs of client requirements. When creating a data-driven Web site, one of the most common tasks Web developers are faced with is creating data entry forms. Data entry forms are Web pages that provide the system's users with a means to input data. The task of creating a particular data entry form typically starts with hammering out the requirements that spell out specifically what information needs to be collected from the user. With the requirements defined, the next stage is designing the data entry Web Form, which involves creating the graphical user interface, as well as writing the code that updates the database with the user's inputs. The main objective of this paper is to construct controls dynamically, that is creating web controls in run time and not in design-time. We can create large amount of dynamic fields with dynamic validations with the help of XML, XSL & Java script. A database plays a major role to accomplish this functionality. We can use 3D (static, dynamic and Meta) database structures. One of the advantages of the XML/XSLT combination is the ability to separate content from presentation. A data source can return an XML document, then by using an XSLT, the data can be transformed into whatever HTML is needed, based on the data in the XML document. The flexibility of XML/XLST can be combined with the power of ASP.NET server/client controls by using an XSLT to generate the server/client controls dynamically, thus leveraging the best of both worlds. This synergy is demonstrated by creating a publication domain application.**

Keywords- three dimensional database, extensible Mark-up Language, web application, dynamic controls, extensible stylesheet language

I. Introduction

When creating a data-driven Web site, one of the most common tasks Web developers are faced with is creating data entry forms. Data entry forms are Web pages that provide the system's users with a means to input data. The task of creating a particular data entry form typically starts with hammering out the requirements that spell out specifically what information needs to be collected from the user. With the requirements defined, the next stage is designing the data entry Web Form, which involves creating the graphical user interface, as well as writing the code that updates the database with the user's inputs.When the data entry forms requirements are well-known in advance, and when such data entry forms are identical across all users for the system, creating such entry forms is hardly challenging. The task becomes more arduous, however, if the data entryforms need to be dynamic. For example, consider a company's Internet Web application whose purpose is tocollect information about the product purchased by a customer; a sort of online product registration system. With such an application, the questions the user is presented with might differ based on what product they purchased, or if they purchased the product from a store or from the company's Web site.When faced with needing to provide dynamic data entry user interfaces, as in the example mentioned above, one option might be to "brute force" a solution. You could create a separate Web page for each product your company sells, with each page having the specific data entry elements needed. The problem with this naive approach is that it requires adding new pages when new products are released. While creating these new pages might not be terribly difficult, it is time consuming and prone to errors without sufficient debugging and testing time.Ideally, when new products are released, a non-technical co-worker could specify what questions are required through an easy-to-use Web-based interface. Such a system is quite possible with ASP.NET thanks to the ability to dynamically load controls on an ASP.NET Web page at runtime. With just a bit of an initial investment in development and testing time, you can create a reusable, dynamic data entry user interface engine. One that allows even the least computer savvy users the ability to easily create customized data entry forms. In this article, we will look at the fundamentals of working with dynamic controls in ASP.NET, and then I will present a complete, working dynamic data entry system that can be easily customized and extended.

II. Existing system

In existing database structure is flat or two dimensional definitions is simple database design consisting of one large table instead of several interconnected tables of a relational database. Called 'flat' because of its only two dimensional (data fields and records) structure, these databases cannot represent complex data relationships. Also called flat file database or flatform database.

Having a flat table to store all the data poses the following issues:

- table grows very large
- indexing the table is problematic
- not optimized for either update or read
- no stringent type checking as everything is stored in the database as a string (varchar/nvarchar)
- catering for text and number is problematic

*About[1]-Computer Science, Karpagam University Pollachi Road, Eachanari, Coimbatore, Tamilnadu India 641 024*vadivel.rangasamy@gmail.com
About[2]-*Computer Science, Karpagam UniversityPollach Road, Eachanari, Coimbatore, Tamilnadu India 641 024*vadivel.rangasamy@gmail.com

- downloading the data requires several joins to group and instance tables which has an impact on performance and adds complexity to the query

A static field structure is a created for an input data set which is not supposed to change within the scope of the problem. When a single field is to be added or deleted, the update of a static field structure incurs significant costs, often comparable with the construction of the field structure from scratch. Create a static field it should know the all required fields and also to take much development time and need to given static names for those fields on design time and also there is no possible to apply unique style. A main disadvantage of static field is not specifying the data types (such as integer, string and etc.,) on run time and flat database structures are used. "N" number of lines to taken creating static fields and occupy the memory on design time and it may be possible to leakage of memory. In hacker can be easily hack those fields on run time and they can create the pseudo code to collapse those fields. Testing results for web controls on one web page, size of web page is 6.0 KB, available static fields is 7 and apply for 5 iterations so that get 35 fields and given below controls loading time for each iterations,

Iterations	Page size (Run Time)	Loading Time
1.	150.2K	2.262s
2.	150.2K	1.258s
3.	150.2K	1.766s
4.	150.2K	1.794s
5.	150.2K	1.484s
Total	751 K	8.564s

III. Related Works

The three dimensional database has been played major rule on the creation dynamic fields. Database architecture for creating dynamic web controls is a three dimensional structure, where we use three terms static, meta and dynamic. Here Static data is generally creating the tables and fields to the database. Meta data is a bridge between static and dynamic data. Dynamic data is the dynamic resultant tables or views that the user needs. An output of database is XML format and it contains data definition and data values. XSL is a presentation part which transforms XML data to output HTML.In three dimensional databases has used two types of SQL statements Static and Dynamic. Static SQL is SQL statements in an application that do not change at runtime and, therefore, can be hard-coded into the application. Dynamic SQL is SQL statements that are constructed at runtime; for example, the application may allow users to enter their own queries. Thus, the SQL statements cannot be hard-coded into the application.XSLT is designed for use as part of XSL, which is a style sheet language for XML. In addition to XSLT, XSL includes an XML vocabulary for specifying formatting. XSL specifies the styling of an XML document by using XSLT to describe how the document is transformed into another XML document that uses the formatting vocabulary.XSLT is also designed to be used independently of XSL. However, XSLT is not intended as a completely general-purpose XML transformation language. Rather it is designed primarily for the kinds of transformations that are needed when XSLT is used as part of XSL.

1) *Data Model*

The data model used by XSLT is the same as that used by XPath with the additions described in this section. XSLT operates on source, result and stylesheet documents using the same data model. Any two XML documents that have the same tree will be treated the same by XSLT. Processing instructions and comments in the stylesheet are ignored: the stylesheet is treated as if neither processing instruction nodes nor comment nodes were included in the tree that represents the stylesheet.

2) *Root Node Children*

The normal restrictions on the children of the root node are relaxed for the result tree. The result tree may have any sequence of nodes as children that would be possible for an element node. In particular, it may have text node children, and any number of element node children. When written out using the XML output method (see [16 Output]), it is possible that a result tree will not be a well-formed XML document; however, it will always be a well-formed external general parsed entity.When the source tree is created by parsing a well-formed XML document, the root node of the source tree will automatically satisfy the normal restrictions of having no text node children and exactly one element child. When the source tree is created in some other way, for example by using the DOM, the usual restrictions are relaxed for the source tree as for the result tree.

3) *Base URI*

Every node also has an associated URI called its base URI, which is used for resolving attribute values that represent relative URIs into absolute URIs. If an element or processing instruction occurs in an external entity, the base URI of that element or processing instruction is the URI of the external entity; otherwise, the base URI is the base URI of the document. The base URI of the document node is the URI of the document entity. The base URI for a text node, a comment node, an attribute node or a namespace node is the base URI of the parent of the node.

IV. Experimental Results

Solis architecture provides three main areas of functionality self-updating interface on the web, robust database administration, searchable front-end for end users. That system is designed so that dynamic data at the core of the integrated system is available in any output or view. The data administrator has control over the data content, various templates and user permissions, thereby giving an unrivalled level of flexibility and control in content collection, management and presentation.

1) *Data Administration*

Acomplete database administration application that provides the full range of control and flexibility needed for complete editorial control over any type of database. Features of the Data Administration component include:

- User permission management
- Saved-search templates to speed data interrogation
- Management control over:
- field types and structure
- data groups
- data viewing tabs
- data record structure
- taxonomy and categorization
- flat downloads
- online subscribers
- Output listing levels.
- Approvals system for self-updated submissions
- Sub-group counting by specification
- Data entity linking

2) *Self-Updating*

This component is an advanced user permission/restriction that enables the data administrator to give access per data record to both editors and directly to users. The access is via an editorial interface that enables the user to access and update information pertaining to a specific data record. Changes and information are submitted into the Approvals are of the Data Administration.

- Fully customizable
- E-commerce capability
- Dynamic paths
- Graphics and rich media upload
- Data record linking (single owner, multiple records)

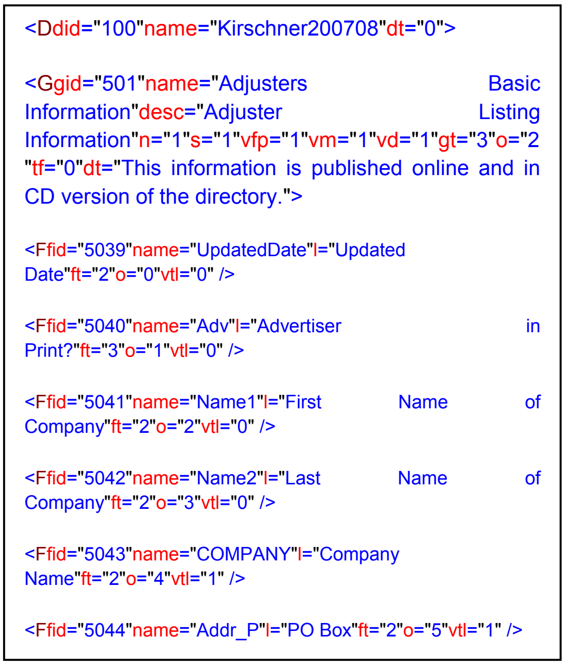
3) *Customer-Facing Front-end*

This component enables the system owner to create custom print or online views of the database that can be integrated into existing websites.

- Web search versions (using template results sets)
- E-commerce control
- Search reconfiguration
- Advertising support
- Permission-based data download
- Automate data output to print (using Adobe InDesign and/or Quark)

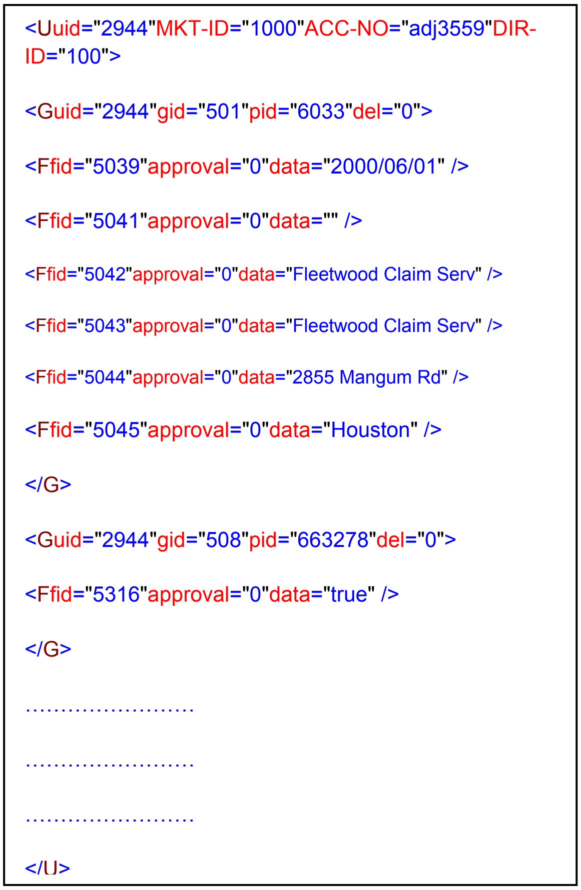
4) *Methodology*

Database: The proposed database architecture for creating dynamic web controls is a three dimensional structure, where we use three terms static, meta and dynamic. Here Static data is generally creating the tables and fields to the database. Meta data is a bridge between static and dynamic data. Dynamic data is the dynamic resultant tables or views that the user needs.XML / XSL: The proposed XML comprises the data definition and data values. Data definition contains a label names and data values contains label values. XSL is a presentation part which transforms XML data to output HTML. Here the screen have show the output of the XML format



```
<Ddid="100"name="Kirschner200708"dt="0">

<Ggid="501"name="Adjusters Basic
Information"desc="Adjuster Listing
Information"n="1"s="1"vfp="1"vm="1"vd="1"gt="3"o="2
"tf="0"dt="This information is published online and in
CD version of the directory.">

<Ffid="5039"name="UpdatedDate"l="Updated
Date"ft="2"o="0"vtl="0" />

<Ffid="5040"name="Adv"l="Advertiser in
Print?"ft="3"o="1"vtl="0" />

<Ffid="5041"name="Name1"l="First Name of
Company"ft="2"o="2"vtl="0" />

<Ffid="5042"name="Name2"l="Last Name of
Company"ft="2"o="3"vtl="0" />

<Ffid="5043"name="COMPANY"l="Company
Name"ft="2"o="4"vtl="1" />

<Ffid="5044"name="Addr_P"l="PO Box"ft="2"o="5"vtl="1" />
```

Fig – 1 XML format for data definition



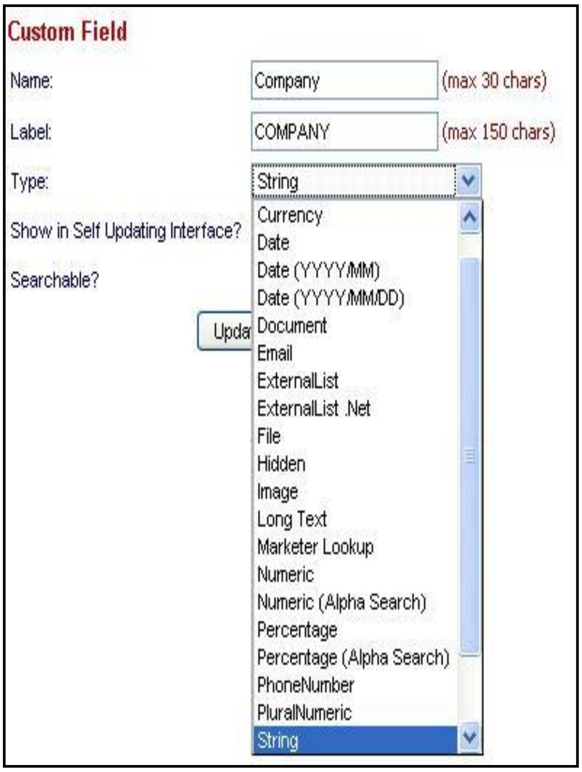
```
<Uuid="2944"MKT-ID="1000"ACC-NO="adj3559"DIR-
ID="100">

<Guid="2944"gid="501"pid="6033"del="0">

<Ffid="5039"approval="0"data="2000/06/01" />

<Ffid="5041"approval="0"data="" />

<Ffid="5042"approval="0"data="Fleetwood Claim Serv" />

<Ffid="5043"approval="0"data="Fleetwood Claim Serv" />

<Ffid="5044"approval="0"data="2855 Mangum Rd" />

<Ffid="5045"approval="0"data="Houston" />

</G>

<Guid="2944"gid="508"pid="663278"del="0">

<Ffid="5316"approval="0"data="true" />

</G>

……………………

……………………

……………………

</U>
```

Fig – 2 XML format for data value

Building blocks of XML documents are nested, tagged elements. Each tagged element has zero or more sub elements; zero or more attribute, and may contain textual information (data content). Elements can be nested at any depth in the document structure. Attributes can be of different types, allowing one to specify element identifiers (attributes of type ID), additional information about the element (e.g., attribute of type CDATA containing textual information), or link to other elements of the document (attributes of type IDREF(s)). An example of XML document is presented in Figure 1 and 2. The document represents the data definition and data values of the publication fields. The XML document contains also all information on the custom fields.To develop on a formal basis our approach for secure publishing of XML documents we introduce a formal model of XML documents that we use throughout the paper. In the following, we denote with Label be a set of element tags and attribute names, and Value a set of attribute/element values. An XML document can be formally defined as follows.

```
<xsl:templatename="DisplayFieldValue">

<xsl:paramname="GroupID" />

<xsl:paramname="InstanceID" />

<xsl:paramname="FieldID" />

<xsl:paramname="FieldType" />

<xsl:paramname="FieldXInfo" />

<xsl:paramname="IsTable" />

<xsl:paramname="ExternalValues" />

<xsl:variablename="dataValue">

<xsl:choose>

<xsl:whentest="/ROOT/U/G[@gid=$GroupID          and
@pid=$InstanceID]/F[@fid=$FieldID and @approval !=
'0']">

<xsl:value-ofselect="/ROOT/U/G[@gid=$GroupID    and
@pid=$InstanceID]/F[@fid=$FieldID and @approval !=
'0']/@data" />

</xsl:when>

<xsl:otherwise>

<xsl:value-ofselect="/ROOT/U/G[@gid=$GroupID    and
@pid=$InstanceID]/F[@fid=$FieldID and @approval =
'0']/@data" />
```

Fig – 3 XSLT for display the XML files into web



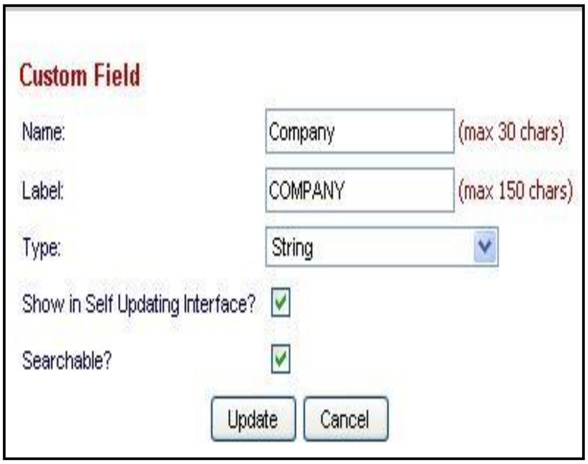
Custom Field

Name: Company (max 30 chars)

Label: COMPANY (max 150 chars)

Type: String

Fig – 4 Display custom data types



Custom Field

Name: Company (max 30 chars)

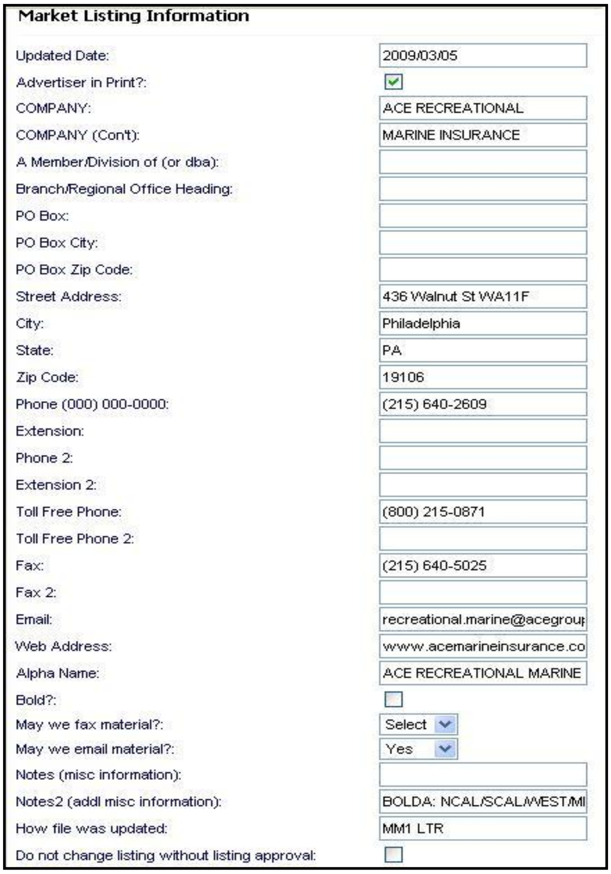
Label: COMPANY (max 150 chars)

Type: String

Show in Self Updating Interface? ☑

Searchable? ☑

Fig – 5 Field creations



Market Listing Information

Updated Date:	2009/03/05
Advertiser in Print?:	☑
COMPANY:	ACE RECREATIONAL
COMPANY (Con't):	MARINE INSURANCE
A Member/Division of (or dba):	
Branch/Regional Office Heading:	
PO Box:	
PO Box City:	
PO Box Zip Code:	
Street Address:	436 Walnut St WA11F
City:	Philadelphia
State:	PA
Zip Code:	19106
Phone (000) 000-0000:	(215) 640-2609
Extension:	
Phone 2:	
Extension 2:	
Toll Free Phone:	(800) 215-0871
Toll Free Phone 2:	
Fax:	(215) 640-5025
Fax 2:	
Email:	recreational.marine@acegrouj
Web Address:	www.acemarineinsurance.co
Alpha Name:	ACE RECREATIONAL MARINE
Bold?:	□
May we fax material?:	Select
May we email material?:	Yes
Notes (misc information):	
Notes2 (addl misc information):	BOLDA: NCAL/SCAL/WEST/MI
How file was updated:	MM1 LTR
Do not change listing without listing approval:	□

Fig – 6 Display field with values

5) *Outputs*

Fig 1 – is a Data Definition XML and herewith details about tags "D" is data, "G" is group, "gid" is a group id, "F" is field, "fid" is field id, "name" is name for the field, "l" is label, "ft" is field data type and "o" is display order.

Fig 2 – is a Data Value XML and herewith details about tags, "U" is a Definition about the data, "G" is group, "gid" is a group id, "F" is field, "fid" is field id, "approval" is a status of the data. There is three types of status that is "0" approved data, "1" data has been newly added or edited existing but waiting for approval and "2" data has been deleted but waiting for approval and "data" is holding on current data.

Fig 3 – Display those XML's to web pages using XSL syntax

Fig 4 – Displaying custom data types which implemented in web application

Fig 5 – Creating field with them data types and custom label for display the user editable field.

Fig 6 – Display the custom fields with values. Here textbox displays when data type is string, dropdown displays when data type is external list and checkbox displays when data type is yes/no.

V. Conclusion

In this article, produced core of the three dimensional database structure and it have five key terms that *directory, entities, groups, fields and field values* and rationalization of those key terms is directory has many entities. Each entity has many groups but entity must have one primary group and implemented successfully on publication domain. Implemented three dimensional databases to product based projects.Web pages consist of a control hierarchy, which is usually composed strictly of statically-defined controls. However, at runtime we can manipulate this control hierarchy by adding dynamic controls to the Controls collection of existing controls in the hierarchy. We also looked at techniques for accessing dynamically-added controls and common patterns for adding and interacting with these controls.Herewith showed statical information about implemented three dimensional database and testing results for dynamic web controls on one web page, size of web page is 10.2 KB, available dynamic fields is more than 10 and apply for 5 iterations so that get more than 50 fields and given below controls loading time for each iterations,

Iterations	Page size (Run Time)	Loading Time
1.	203.7K	1.602s
2.	203.7K	1.486s
3.	203.7K	1.430s
4.	203.7K	1.540s
5.	203.7K	1.270s
Total	1018.5K	7.328s

Being able to manipulate a web page's control hierarchy at runtime is a powerful and useful tool that has applications in many common scenarios. Armed with this article, you should be able to confidently work with dynamic controls in your web pages.

VI. Reference

1) Ke Yi , Feifei Li , Graham Cormode , Marios Hadjieleftheriou , George Kollios , Divesh Srivastava, Small synopses for group-by query verification on outsourced data streams, ACM Transactions on Database Systems (TODS), v.34 n.3, p.1-42, August 2009

2) HweeHwa Pang , Jilian Zhang , Kyriakos Mouratidis, Scalable verification for outsourced dynamic databases, Proceedings of the VLDB Endowment, v.2 n.1, August 2009

3) Alberto Trombetta, Danilo Montesi, "Equivalences and Optimizations in an Expressive XSLT Fragment" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 11, JULY 2009

4) Kyriakos Mouratidis , Dimitris Sacharidis , Hweehwa Pang, Partially materialized digest scheme: an efficient verification method for outsourced databases, The VLDB Journal — The International Journal on Very Large Data Bases, v.18 n.1, p.363-381, January 2009

5) HweeHwa Pang , Kyriakos Mouratidis, Authenticating the query results of text search engines, Proceedings of the VLDB Endowment, v.1 n.1, August 2008
6) http://www.dnzone.com/go?151&LinkFile=page1.htm
7) http://msdn.microsoft.com/en-us/library/aa479330.aspx

Improving Software Effort Estimation Using Neuro-Fuzzy Model with SEER-SEM

Wei Lin Du[1], Danny Ho[2], Luiz Fernando Capretz [3]

GJCST Classification
D.2.9, K.6.3, K.6.4

Abstract **- Accurate software development effort estimation is a critical part of software projects. Effective development of software is based on accurate effort estimation. Although many techniques and algorithmic models have been developed and implemented by practitioners, accurate software development effort prediction is still a challenging endeavor in the field of software engineering, especially in handling uncertain and imprecise inputs and collinear characteristics. In order to address these issues, previous researchers developed and evaluated a novel soft computing framework. The aims of our research are to evaluate the prediction performance of the proposed neuro-fuzzy model with System Evaluation and Estimation of Resource Software Estimation Model (SEER-SEM) in software estimation practices and to apply the proposed architecture that combines the neuro-fuzzy technique with different algorithmic models. In this paper, an approach combining the neuro-fuzzy technique and the SEER-SEM effort estimation algorithm is described. This proposed model possesses positive characteristics such as learning ability, decreased sensitivity, effective generalization, and knowledge integration for introducing the neuro-fuzzy technique. Moreover, continuous rating values and linguistic values can be inputs of the proposed model for avoiding the large estimation deviation among similar projects. The performance of the proposed model is accessed by designing and conducting evaluation with published projects and industrial data. The evaluation results indicate that estimation with our proposed neuro-fuzzy model containing SEER-SEM is improved in comparison with the estimation results that only use SEER-SEM algorithm. At the same time, the results of this research also demonstrate that the general neuro-fuzzy framework can function with various algorithmic models for improving the performance of software effort estimation.**

Keywords – software estimation, software management, software effort estimation, neuro-fuzzy software estimation, SEER-SEM

I. Introduction

The cost and delivery of software projects and the quality of products are affected by the accuracy of software effort estimation. In general, software effort estimation techniques can be subdivided into experience-based, parametric model-based, learning-oriented, dynamics-based, regression-based, and composite techniques (Boehm, Abts, and Chulani 2000). Amongst these methods, model-based estimation techniques involve the use of mathematical equations to perform software estimation. The estimation effort is a function of the number of variables, which are factors impacting software cost (Boehm 1981). These model-based estimation techniques comprise the general form: $E = a \times Size^b$, where E is the effort, size is the product size, a is the productivity parameters or factors, and b is the parameters for economies or diseconomies (Fischman, McRitchie, and Galorath 2005; Jensen, Putnam, and Roetzheim 2006). In the past decades, some important software estimation algorithmic models have been published by researchers, for instance Constructive Cost Model (COCOMO) (Boehm et al. 2000), Software Life-cycle Management (SLIM) (Putnam and Myers 1992), SEER-SEM (Galorath and Evans 2006), and Function Points (Albrecht 1979; Jones 1998). Model-based techniques have several strengths, the most prominent of which are objectivity, repeatability, the presence of supporting sensitivity analysis, and the ability to calibrate to previous experience (Boehm 1981). On the other hand, these models also have some disadvantages. One of the disadvantages of algorithmic models is their lack of flexibility in adapting to new circumstances. The new development environment usually entails a unique situation, resulting in imprecise inputs for estimation by an algorithmic model. As a rapidly changing business, the software industry often faces the issue of instability and hence algorithmic models can be quickly outdated. The outputs of algorithmic models are based on the inputs of size and the ratings of factors or variables (Boehm 1981). Hence, incorrect inputs to such models, resulting from outdated information, cause the estimation to be inaccurate. Another drawback of algorithmic models is the strong collinearity among parameters and the complex non-linear relationships between the outputs and the contributing factors.

SEER-SEM appeals to software practitioners because of its powerful estimation features. It has been developed with a combination of estimation functions for performing various estimations. Created specifically for software effort estimation, the SEER-SEM model was influenced by the frameworks of Putnam (Putnam and Myers 1992) and Doty Associates (Jensen, Putnam, and Roetzheim 2006). As one of the algorithmic estimation models, SEER-SEM has two main limitations on effort estimation. First, there are over 50 input parameters related to the various factors of a project, which increases the complexity of SEER-SEM, especially for managing the uncertainty from these outputs. Second, the specific details of SEER-SEM increase the difficulty of discovering the nonlinear relationship between the

About[1]-Wei Lin Du, the Department of Electrical and Computer Engineering, the University of Western Ontario, London, Ontario, Canada N6A 5B9(email: wdu6@uwo.ca)

About[2] -Danny Ho, NFA Estimation Inc., Richmond Hill, Ontario Canada L4C 0A2(email: danny@nfa-estimation.com)

About[3]-Dr. Luiz Fernando Capretz, the Department of Electrical and Computer Engineering, the University of Western Ontario, London, Ontario, Canada N6A 5B9 (telephone: 1-519-661-2111 ext. 85482 email: lcapretz@eng.uwo.ca)

parameter inputs and the corresponding outputs. Overall, these two major limitations can lead to a lower accuracy in effort estimation by SEER-SEM.

The estimation effort is a function of the number of variables, which are factors impacting software cost (Boehm 1981). These model-based estimation techniques comprise the general form: E = a × Size^b, where E is the effort, size is the product size, a is the productivity parameters or factors, and b is the parameters for economies or diseconomies (Fischman, McRitchie, and Galorath 2005; Jensen, Putnam, and Roetzheim 2006). In the past decades, some important software estimation algorithmic models have been published by researchers, for instance Constructive Cost Model (COCOMO) (Boehm et al. 2000), Software Life-cycle Model (SLIM) (Putnam and Myers 1992), SEER-SEM (Galorath and Evans 2006), and Function Points (Albrecht 1979; Jones 1998). Model-based techniques have several strengths, the most prominent of which are objectivity, repeatability, the presence of supporting sensitivity analysis, and the ability to calibrate to previous experience (Boehm 1981). On the other hand, these models also have some disadvantages. One of the disadvantages of algorithmic models is their lack of flexibility in adapting to new circumstances. The new development environment usually entails a unique situation, resulting in imprecise inputs for estimation by an algorithmic model. As a rapidly changing business, the software industry often faces the issue of instability and hence algorithmic models can be quickly outdated. The outputs of algorithmic models are based on the inputs of size and the ratings of factors or variables (Boehm 1981). Hence, incorrect inputs to such models, resulting from outdated information, cause the estimation to be inaccurate. Another drawback of algorithmic models is the strong collinearity among parameters and the complex non-linear relationships between the outputs and the contributing factors.

SEER-SEM appeals to software practitioners because of its powerful estimation features. It has been developed with a combination of estimation functions for performing various estimations. Created specifically for software effort estimation, the SEER-SEM model was influenced by the frameworks of Putnam (Putnam and Myers 1992) and Doty Associates (Jensen, Putnam, and Roetzheim 2006). As one of the algorithmic estimation models, SEER-SEM has two main limitations on effort estimation. First, there are over 50 input parameters related to the various factors of a project, which increases the complexity of SEER-SEM, especially for managing the uncertainty from these outputs. Second, the specific details of SEER-SEM increase the difficulty of discovering the nonlinear relationship between the parameter inputs and the corresponding outputs. Our study attempts to reduce the negative impacts of the above major limitations of the SEER-SEM effort estimation model on prediction accuracy and make contributions towards resolving the problems caused by the disadvantages of algorithmic models. First, for accurately estimating software effort the neural network and fuzzy logic approaches are adopted to create a neuro-fuzzy model, which is subsequently combined with SEER-SEM. The Adaptive Neuro-Fuzzy Inference System (ANFIS) is used as the architecture of each neuro-fuzzy sub-model. Second, this research is another evaluation for effectiveness of the general model of neuro-fuzzy with algorithmic model proposed by the previous studies. Third, the published data and industrial project data are used to evaluate the proposed neuro-fuzzy model with SEER-SEM. Although the data was collected specifically for COCOMO 81 and COCOMO 87, they are transferred from COCOMOs to COCOMO II and then to the SEER-SEM parameter inputs, utilizing the guidelines from the University of Southern California (USC) (Madachy, Boehm, and Wu 2006; USC Center for Software Engineering 2006). After the transfer of this data, the estimation performance is verified to ensure its feasibility.

II. Background

Soft computing, which is motivated by the characteristics of human reasoning, has been widely known and utilized since the 1960s. The overall objective from this field is to achieve the tolerance of incompleteness and to make decisions under imprecision, uncertainty, and fuzziness (Nauck, Klawonn, and Kruse 1997; Nguyen, Prasad, Walker, and Walker 2003). Because of capabilities, soft computing has been adopted by many fields, including engineering, manufacturing, science, medicine, and business. The two most prominent techniques of soft computing are neural networks and fuzzy systems. The most attractive advantage of neural networks is the ability to learn from previous examples, but it is difficult to prove that neural networks are working as expected. Neural networks are like ―black boxes" to the extent that the method for obtaining the outputs is not revealed to the users (Chulani 1999; Jang, Sun, and Mizutani 1997). The obvious advantages of fuzzy logic are easy to define and understand an intuitive model by using linguistic mappings and handle imprecise information (Gray and MacDonell 1997; Jang, Sun, and Mizutani 1997). On the other hand, the drawback of this technique is that it is not easy to guarantee that a fuzzy system with a substantial number of complex rules will have a proper degree of meaningfulness (Gray and MacDonell 1997). In addition, the structure of fuzzy if-then rules lacks the adaptability to handle external changes (Jang, Sun, and Mizutani 1997). Although neural networks and fuzzy logic have obvious strengths as independent systems, their disadvantages have prompted researchers to develop a hybrid neuro-fuzzy system that minimizes these limitations. Specifically, a neuro-fuzzy system is a fuzzy system that is trained by a learning algorithm derived from the neural network theory (Nauck, Klawonn, and Kruse 1997). Jang‗s (Jang, Sun, and Mizutani 1997; Nauck, Klawonn, and Kruse 1997) ANFIS is one type of hybrid neuro-fuzzy system, which is composed of a five-layer feed-forward network architecture.

Soft computing is especially important in software cost estimation, particularly when dealing with uncertainty and with complex relationships between inputs and outputs. In the 1990‗s a soft computing technique was introduced to build software estimation models and improve prediction performance (Damiani, Jain, and Madravio 2004). As a technique containing the advantages of the neural networks and fuzzy logic, the neuro-fuzzy model was adopted for software estimation. Researchers developed some models with the neuro-fuzzy technique and demonstrated their ability to improve prediction accuracy. Hodgkinson and Garratt (*Hodgkinson and Garratt 1999)* introduced the neuro-fuzzy model for cost estimation as one of the important methodologies for developing non-algorithmic models. Their model did not use any of the existing prediction models, as the inputs are size and duration, and the output is the estimated project effort. The clear relationship between Function Points Analysis (FPA)'s primary component and effort was demonstrates by Abran and Robillard's study (Abran and Robillard 1996). Huang *et al.* (Huang, Ho, Ren, and Capretz 2005 and 2006) proposed a software effort estimation model that combines a neuro-fuzzy framework with COCOMO II. The parameter values of COCOMO II were calibrated by the neuro-fuzzy technique in order to improve its prediction accuracy. This study demonstrated that the neuro-fuzzy technique was capable of integrating numerical data and expert knowledge. And the performance of PRED(20%) and PRED(30%) were improved by more than 15% and 11% in comparison with that of COCOMO 81. Xia *et al.* (Xia, Capretz, Ho, and Ahmed 2008) developed a Function Point (FP) calibration model with the neuro-fuzzy technique, which is known as the Neuro-Fuzzy Function Point (NFFP) model. The objectives of this model are to improve the FP complexity weight systems by fuzzy logic, to calibrate the weight values of the unadjusted FP through the neural network, and to produce a calibrated FP count for more accurate measurements. Overall, the evaluation results demonstrated that the average improvement for software effort estimation accuracy is 22%. Wong *et al.* (Wong, Ho, and Capretz 2008) introduced a combination of neural networks and fuzzy logic to improve the accuracy of backfiring size estimates. In this case, the neuro-fuzzy approach was used to calibrate the conversion ratios with the objective of reducing the margin of error. The study compared the calibrated prediction model against the default conversion ratios. As a result, the calibrated ratios still presented the inverse curve relationship between the programming languages level and the SLOC/FP, and the accuracy of the size estimation experienced a small degree of improvement.

III. A Neuro-Fuzzy SEER-SEM Model

A. *A General Soft Computing Framework for Software Estimation*

This section describes a general soft computing framework for software estimation, which is based on the unique architecture of the neuro-fuzzy model described in the patent US-7328202-B2 (Huang, Ho, Ren, and Capretz 2008) and was built by Huang *et al.* (Huang, Ho, Ren, and Capretz 2006). The framework is composed of inputs, a neuro-fuzzy bank, corresponding values of inputs, an algorithmic model, and outputs for effort estimation, as depicted in Fig. 1. Among the components of the proposed framework, the neuro-fuzzy bank and the algorithmic model are the major parts of the model. The inputs are rating levels, which can be continuous values or linguistic terms such as Low, Nominal, or High. *V1, …,Vn* are the non-rated values of the software estimation algorithmic model. On the other hand, AI_0, …, AI_m are the corresponding adjusted quantitative parameter values of the rating inputs, which are the inputs of the software estimation algorithmic model for estimating effort as the final output.

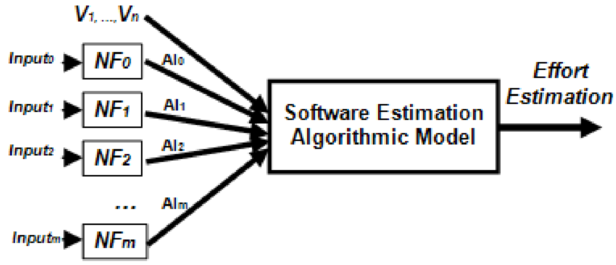


Fig.1. A General Soft Computing Framework.

This novel framework has attractive attributes, particularly the fact that it can be generalized to many different situations and can be used to create more specific models. In fact, its generalization is one of the purposes of designing this framework. Its implementation is not limited to any specific software estimation algorithmic model. The algorithmic model in the framework can be one of the current popular algorithmic models such as COCOMO, SLIM or SEER-SEM. When various algorithmic models are implemented into this framework, the inputs and the non-rating values are different.

B. *SEER-SEM Effort Estimation Model*

SEER-SEM stemmed from the Jensen software model in the late 1970s, where it was developed at the Hughes Aircraft Company's Space and Communications Group (Fischman, McRitchie, and Galorath 2005; Galorath and Evans 2006; Jensen, Putnam, and Roetzheim 2006). In 1988, Galorath Inc. (GAI) started developing SEER-SEM (Galorath and Evans 2006), and in 1990, GAI trademarked this model. The SEER-SEM model was motivated by Putnam's SLIM and Boehm's COCOMO (Fischman, McRitchie, and Galorath 2005; Galorath and Evans 2006; Jensen, Putnam, and Roetzheim 2006). Over the span of a decade, SEER-SEM has been developed into a powerful and sophisticated model, which contains a variety of tools for performing different estimations that are not limited to software effort. SEER-SEM includes the breakdown structures for various tasks, project life cycles, platforms, and applications. It also includes the most development languages, such as the third and fourth generation programming languages, in the estimation. Furthermore, the users can select different knowledge bases (KBs) for Platform, Application, Acquisition Method, Development Method, Development Standard, and Class based on the requirements of their projects. SEER-SEM provides the baseline settings for parameters according to the KB inputs; there are over 50 parameters that impact the estimation outputs. Among them, 34 parameters are used by SEER-SEM effort estimation model (Galorath Incorporated 2001 and 2006). Nevertheless, the SEER-SEM model contains some disadvantages. For instance, the efforts spent on pre-specification phases, such as requirements collection, are not included in the effort estimation. In SEER-SEM effort estimation, each parameter has sensitivity inputs, with the ratings ranging from Very Low (VLo-) to Extra High (EHi+). Each main rating level is divided into three sub-ratings, such as VLo-, VLo, VLo+. These ratings are translated to the corresponding quantitative value used by the effort estimation calculation. The SEER-SEM effort estimation is calculated by the following equations:

$$E = 0.393469 \times K \qquad (1)$$

$$C_{tb} = 2000 \times \exp\left(\frac{-3.70945 \times \ln\left(\frac{ctbx}{4.11}\right)}{5 \times TURN} \right) \qquad (2)$$

$$K = D^{0.4} \times \left(\frac{S_e}{C_{te}}\right)^{1.2}, C_{te} = C_{tb} / ParmAdjustment \qquad (3)$$

ctbx =
ACAP × AEXPAPPL × MODP × PCAP × TOOL × TERM (4)

ParmAdjustment=
LANGLEXP × TSYSTEXP × DSYSDEXP × PSYSPEXP × SIBRREUS × MULT × RDED × RLOC × DSVL × PSVL × RVOL × SPEC × TEST × QUAL × RHST(HOST) × DISP × MEMC × TIMC × RTIM × SECR × TSVL (5)

where,
E is the development effort (in person years),
K is the total Life-cycle effort (in person years) including development and maintenance,
S_e is the Effective Size (SLOC),
D is the Staffing complexity,
C_{te} is the Effective technology,
C_{tb} is the Basic technology.

The elements included in equations (4) and (5) are parameters or combined parameters; the formulas for calculating combined parameters are shown below:

AEXPAPPL =
0.82+(0.47*EXP(-0.95977*(AEXP/APPL))) (6)

LANGLEXP =
1+((1.11+0.085*LANG)-1)*EXP(-LEXP/(LANG/3)) (7)
TSYSTEXP =
1+(0.035+0.025*TSYS)*EXP(-3*TEXP/TSYS) (8)
DSYSDEXP =
1+(0.06+0.05*DSYS)*EXP(-3*DEXP/DSYS) (9)

PSYSPEXP

$$\begin{cases} = (0.91\text{^ PSYS} + 0.23 * \text{PSYS} * \text{EXP(-3} * \text{PEXP /PSYS))\^{}0 .833,} \\ \quad \textit{when}\ \text{PSYS} \neq 0 \\ = 1, \textit{when}\ PSYS = 0 \end{cases} \qquad (10)$$

SIBRREUS =
SIBR*REUS +1 (11)

C. *A Neuro-Fuzzy Model with SEER-SEM*

a) *Overview*

This section will describe the proposed framework of the neuro-fuzzy model with SEER-SEM, based on the general structure in the section III.A, as depicted in Fig. 2. The inputs consist of two parts: non-rating inputs and the rating levels of parameters, which include 34 technology and environment parameters and 1 complexity or staffing parameter. Among the technology and environment parameters, there is one parameter (SIBR), which is not rated by the linguistic term. SIBR is decided by users, through inputting the percentage. Hence, similar to the input of size, SIBR is a non-rating value. While the other parameters are labeled as PR_1 *to* PR_{34}, SIBR is labeled PR_{35}.

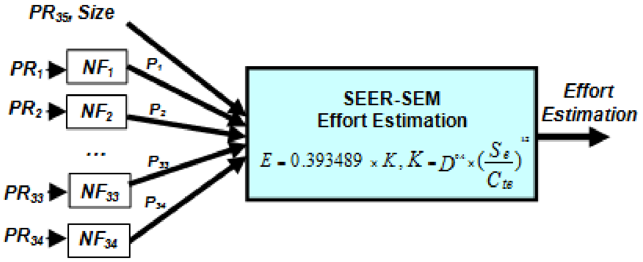


Fig.2. A Neuro-Fuzzy Model with SEER-SEM.

Each parameter PRi (i = 1, …, 34) can be a linguistic term or a continuous rating value. The linguistic inputs are from 18 rating levels (r =1, …, 18), which include Very Low– (VLo-), Very Low (VLo), Very Low+ (VLo+), Low–, Low, Low+, Nominal- (Nom-), Nominal (Nom), Nominal+ (Nom+), High – (Hi-), High (Hi), High+ (Hi+), Very High– (VHi-), Very High (VHi), Very High+ (VHi+), Extra High– (EHi-), Extra High (EHi), and Extra High+ (EHi+). In these ratings, there are 6 main levels, VLo, Low, Nom, Hi, VHi, and EHi, and each main rating level has three sub-levels: minus, plus or neutral (Galorath Incorporated 2006 be 2005). NFi (i = 1, …, 34) is a neuro-fuzzy bank, which is composed of thirty-four NFi sub-models. The rating levels of each parameter PRi (i = 1, …, 34) are the input of each NFi . Through these sub-models, the rating level of a parameter is translated into the corresponding quantitative value (Pi , i = 1, …, 34) as the inputs of the SEER-SEM effort estimation as introduced in the section III.B, from equations (1) to (11). The output of the proposed model is the software effort estimation.

b) Structure of NF_i

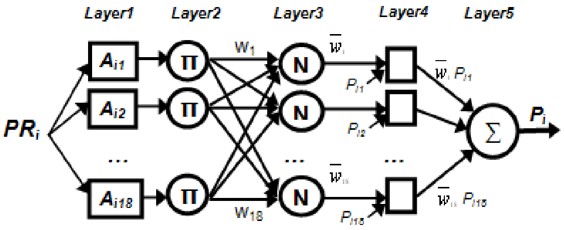


Fig.3. Structure of *NF_i*.

The neuro-fuzzy bank fulfills an important function in the proposed neuro-fuzzy model with SEER-SEM effort estimation model. *NF_i* produces fuzzy sets and rules for training datasets. It translates the rating levels of a parameter into a quantitative value and calibrates the value by using actual project data. According to fuzzy logic techniques, linguistic terms can be presented as a fuzzy set. There are 18 rating levels for each parameter in linguistic terms, which are used to define a fuzzy set in this research. The selected membership function translates the linguistic terms in this fuzzy set to membership values. Each *NF_i* uses the structure of the Adaptive Neuro-Fuzzy Inference System (ANFIS), which is a five-layer hybrid neuro-fuzzy system, as depicted in Fig. 3.

- **Input and Output of *NF_i***

There is one input and one corresponding output for each NF. The input of each *NF_i* (*PR_i* , *i* = 1, …, 34) is the rating level of a parameter for SEER-SEM effort estimation model, such as Very Low (VLo) or High (Hi). On the other hand, the output is the corresponding quantitative value of this parameter (*P_i* , *i* = 1, …, 34), such as 1.30.

- **Fuzzy Rule**

Based on the features of ANFIS and the structure shown in Fig. 3, this work refers to the form of the fuzzy if-then rule proposed by Takagi and Sugeno (Takagi and Sugeno 1986). The *r*th fuzzy rule of the proposed model is defined as below:

Fuzzy Rule r: **IF** *PR_i* is *A_{ir}* **THEN** *$P_i = P_{ir}$* ,
r =1, 2, …, 18

where *A_{ir}* is a rating level of the fuzzy set that ranges from Very Low- to Extra High+ for the *i*th parameter and is characterized by the selected membership function, and *P_{ir}* is the corresponding quantitative value of the *r*th rating level for the *i*th parameter. Furthermore, with this fuzzy rule, the premise part is the fuzzy set and the consequent part is the non-fuzzy value. Overall, the fuzzy rules build the links between a linguistic rating level and the corresponding quantitative value of a parameter.

- **Functions of Each Layer**

***Layer 1*:** In this layer, the membership function of fuzzy set *A* translates the input, *PR_i*, to the membership grade. The output of this layer is the membership grade of *PR_i* , which is the premise part of fuzzy rules. Also, the membership function of the nodes in this layer is utilized as the activation function; in our proposed model, all the membership functions of each node in Layer 1 are the same. In subsequent sections, the selected membership function will be discussed in detail.

$$O_r^1 = \mu_{A_{ir}}(PR_i) \quad \text{for } i = 1, 2, \ldots, 34 \quad r = 1, 2, \ldots, 18 \qquad (12)$$

where O_k^i is the membership grade of *A_{ir}* (=VLo-, VLo, VLo+, Low-, Low, Low+, Nom-, Nom, Nom+, Hi-, Hi, Hi+, VHi-, VHi, VHi+, EHi-, EHi, or EHi+) with the input *PR_i* or $\mu_{A_{ir}}$ continuous number $x \in [0,19]$; is the membership function of *A_{ir}* .

***Layer 2*:** Producing the firing strength is the primary function of this layer. The outputs of Layer 1 are the inputs of each node in this layer. In each node, Label Π multiplies all inputs to produce the outputs according to the defined fuzzy rule for this node. Consequently, the outputs of this layer are the firing strength of a rule. The premise part in the defined fuzzy rule of our proposed model is only based on one condition. Therefore, the output of this layer, the firing strength, is not changed and is thus the same as the inputs, or membership grade.

$$O_r^2 = w_r = O_r^1 = \mu_{A_{ir}}(PR_i) \qquad (13)$$

Layer 3: The function of this layer is to normalize the firing strengths for each node. For each node, labeled "N", the ratio of the rth rule's firing strength to the sum of all rules' firing strengths related to PRi is calculated. The resulting outputs are known as normalized firing strengths.

$$O_r^3 = \overline{w_r} = \frac{w_r}{\sum_{r=1}^{18} w_r} \qquad (14)$$

Layer 4: An adaptive result of *P_i* is calculated with the Layer 3 outputs and the original input of *P_i* in the fuzzy rules by multiplying $\overline{w_r}$. The outputs are referred to as consequent parameters.

$$O_r^4 = \overline{w_r}\, P_{ir} \qquad (15)$$

Layer 5: This layer aims to compute the overall output with the sum of all reasoning results from Layer 4.

$$O_r^5 = \sum_r O_r^4 = P_i = \sum_r \overline{w_r}\, P_{ir} \qquad (16)$$

- **Membership Function**

This section describes the triangular membership function utilized in this work; this particular function is depicted in Fig. 4. Each rating level has the corresponding triangular membership function. This membership function is a piecewise-linear function. Throughout the learning process, the membership function is maintained in a fixed state. The following calculation defines the triangular membership function:

$$\mu_{A_{ir}}(x)=\begin{cases} x-(r-1), r-1\le x\le r \\ (r+1)-x, r\le x\le r+1 \\ 0, otherwise \end{cases} \text{ for } r = 1, 2, \ldots, 18 \tag{17}$$

$$x\in[0,19]$$

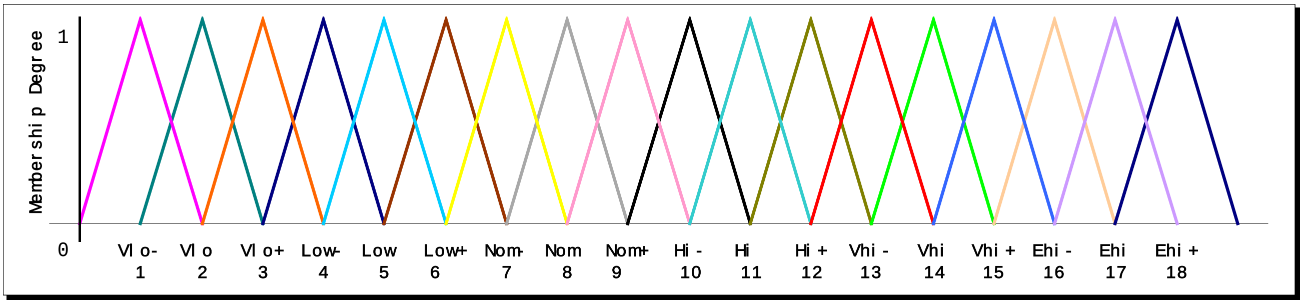


Fig.4. Triangular Membership Function

There are several factors that influenced our selection of the triangular membership function; first, the nature of the NFi outputs was the most crucial reason. Pir is a piecewise-linear interpolation

$$\frac{y-y_0}{y_1-y_0}=\frac{x-x_0}{x_1-x_0}$$

between parameter values ($P_{i1}, \ldots P_{i18}$) of the ith parameter, $P_{i.}$ Hence, the selection of the triangular function can be derived from the same results as a linear interpolation. Secondly, one of the purposes of this research is to evaluate the extent to which Huang's proposed soft computing framework can be generalized. Therefore, it was important to use the same membership function as that utilized in Huang's research in order to perform validation with a similar fuzzy logic technique (Huang 2003). Finally, the triangular membership function is easy to calculate.

- **Learning Algorithm**

With ANFIS, there is a two-pass learning cycle: the forward pass and the backward pass. The pass that is selected depends on the trained parameters in ANFIS. In our proposed model, when the error exists between the actual effort and the estimated effort, the outputs are fixed and the inputs are trained. Hence, the backward pass is the type of learning algorithm that this study uses. It is generally a hybrid method of Least Square Estimate (LSE) and Back Propagation, which is calculated using a gradient decent algorithm that minimizes the error. For the learning algorithm, the parameters of the premise part and the consequent part are defined in two sets, as illustrated below:

$$X = \{x_1, x_2, \ldots, xn\}$$
$$= \{PR_1, PR_2, \ldots, PR_N, SIBR, \text{Size}\} \tag{18}$$
$$P = \{\{P_{11}, P_{21}, \ldots, P_{N1}\}, \{P_{12}, P_{22}, \ldots, P_{N2}\}, \ldots, \{P_{1M}, P_{2M}, \ldots, P_{NM}\}\} \tag{19}$$

where N = 34 and M =18; X represents the inputs of the model, which are the rating levels, SIBR and Size; and P is the parameter values of the parameters.

The output of each NF can be defined when substituting (13) and (14) into (16):

$$P_i = f_{NF_i}(P_{i1,} P_{i2},\ldots, P_{i18}) = \sum_r \overline{w_r} P_{ir} = \sum_{r=1}^{18} \mu_{A_{ir}}(x_i) P_{ir}$$
for i = 1, 2, …, 34 (20)

P_i is the weighted sum of inputs X for PR_i .

In the section III.B, the equations for the SEER-SEM Effort Estimation are described in detail. The equations (1), (2), (3), (4), and (5) can be re-written as follows with the parameters symbols:

$$Effort = 0.393469 \times P_{34}^{0.4} \times \frac{Size^{1.2}}{2000^{1.2} \times \exp\left(\frac{-3.70945\times\ln\left(\frac{ctbx}{4.11}\right)}{5\times P_{10}}\right)^{1.2}} \times ParmAdjustment^{1.2} \tag{21}$$

$$ctbx = P_1 \times P_{2\text{-}25} \times P_8 \times P_3 \times P_9 \times P_{11} \tag{22}$$

$$ParmAdjustment = P_{23\text{-}4} \times P_{31\text{-}6} \times P_{24\text{-}5} \times P_{26\text{-}7} \times P_{35\text{-}22} \times P_{12} \times \ldots \times P_{21} \times P_{27} \times \ldots \times P_{30} \times P_{33} \times P_{32} \tag{23}$$

Utilizing equations (18) to (21), the proposed neuro-fuzzy model can be written:

$$Effort = f_{NF}(X, P) \qquad (24)$$

If there are NN project data points, the inputs and outputs can be presented as (X_n, E_{acn}), where n = 1, 2,…, NN, X_n contains 34 parameters as well as SIBR and Size, E_{aen} is the actual effort with X_n inputs for project n. The learning procedure involves adopting the gradient descent method to adjust the parameter values of rating levels that minimizes the error, *E*. According to LSE, the error, *E*, on the output layer is defined as follows:

$$E = \frac{1}{2}\sum_{n=1}^{NN} w_n \left(\frac{E_{en} - E_{acn}}{E_{acn}}\right)^2 \qquad (25)$$

where w_n is the weight of project n and E_{en} is the estimation of the output for project n.

$$E_{en} = Effort_n = f_{NF}(X_n, P_n) \qquad (26)$$

The following steps are used to perform gradient descent according to the Back Propagation learning algorithm. According to the SEER-SEM effort estimation model presented by equations (21) to (23), the results of the partial derivative of E_{en} with respect to

P_{ir}, $\frac{\partial E_{en}}{\partial P_{ir}}$, are different.

$$\frac{\partial E}{\partial P_{ir}} = \sum_{n=1}^{NN} \frac{w_n}{E_{en}^2}\left(E_{en} - E_{acn}\right)\frac{\partial E_{en}}{\partial P_{ir}} \qquad (27)$$

$$\frac{\partial E_{en}}{\partial P_{ir}} = \frac{\partial E_{en}}{\partial P_i}\frac{\partial P_i}{\partial P_{ir}} = \frac{\partial(f_{NF}(X_{n,}P_n))}{\partial P_i}\frac{\partial P_i}{\partial P_{ir}}$$

for *i* = 1, 2, …, 34 (28)

$$\frac{\partial P_i}{\partial P_{ir}} = \frac{\partial(f_{NF_i}(P_{ir}))}{\partial P_{ir}} = \frac{\partial(\mu_{A_{ir}}(x_{ir})P_{ir})}{\partial P_{ir}} = \mu_{A_{ir}}(x_i) \qquad (29)$$

$$\frac{\partial E_{en}}{\partial P_{ir}}$$

After is calculated out, equation (30) is used to calculate the adjusted parameter values.

$$P_{ir}^{l+1} = P_{ir}^{l} - \alpha\frac{\partial E}{\partial P_{ir}} \qquad (30)$$

where *α* > 0 is the learning rate and *l* is the current iteration index.

- **Monotonic Constraints**

A monotonic function is a function that preserves the given order. The parameter values of SEER-SEM are either monotonic increasing or monotonic decreasing. The relationship between the monotonic functions and the rating levels have been accepted by the practitioners as a common sense practice. For instance, the values of ACAP are monotonic decreasing from VLo- to EHi+, which is reasonable because the higher the analysts' capability, the less spent on project efforts. As for TEST, its values are monotonic increasing because the higher test level causes more effort to be spent on projects. After calibrating parameter values by the proposed model, the trained results of these values may contravene the monotonic orders, so that the trained values are changed to a non-monotonic order. For instance, the parameter value of the ACAP rating Hi can be greater than the value of the corresponding rating, EHi. This discrepancy can lead to unreasonable inputs for performing estimation and can impact the overall accuracy. Therefore, monotonic constraints are used by our model in order to maintain consistency with the rating levels.

IV. Evaluation

For evaluating the neuro-fuzzy SEER-SEM model, in total, data from 99 studies is collected, including 93 published COCOMO 81 projects and 6 industry studies in the format of COCOMO 87 (Ho 1996; Panlilio-Yap and Ho 2000). An algorithmic estimation model, E = a×Size^b comprises the general form of COCOMO and SEER-SEM (Fischman, McRitchie, and Galorath 2005; Jensen, Putnam, and Roetzheim 2006). Specifically, this model enables us to use the COCOMO database for evaluating the proposed SEER-SEM model in spite of the difference between COCOMO and SEER-SEM. In fact, various studies have revealed the similar estimation performances of COCOMO and SEER-SEM (Madachy, Boehm, and Wu 2006; USC Center for Software Engineering 2006).

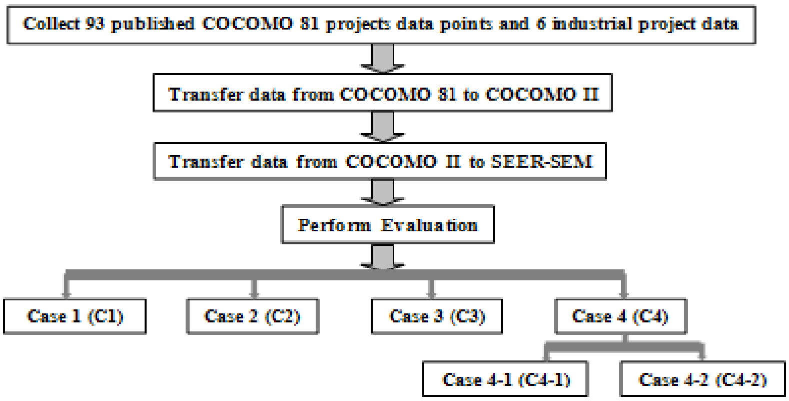

Collect 93 published COCOMO 81 projects data points and 6 industrial project data
Transfer data from COCOMO 81 to COCOMO II
Transfer data from COCOMO II to SEER-SEM
Perform Evaluation
Case 1 (C1)
Case 2 (C2)
Case 3 (C3)
Case 4 (C4)
Case 4-1 (C4-1)
Case 4-2 (C4-2)

Fig.5. Main Evaluation Steps.

Fig. 5 shows the main steps of our evaluation. First, in order to use both published COCOMO 81 and industrial project data in the evaluation, the information was translated into the corresponding format of SEER-SEM data. Second, there are four cases for evaluating the prediction performance of our neuro-fuzzy model.

1) *Performance Evaluation Metrics*

The following evaluation metrics are adapted to assess and evaluate the performance of the effort estimation models.

- **Relative Error (RE)**

$$RE = \frac{(EstimationEffort - ActualEffort)}{ActualEffort}$$

The RE is used to calculate the estimation accuracy.

- **Magnitude of Relative Error (MRE)**

$$MRE = \frac{|EstimationEffort - ActualEffort|}{ActualEffort}$$

- **Mean Magnitude of Relative Error (MMRE)**

$$MMRE = \frac{\left(\sum_{i=1}^{n} MRE_i\right)}{n}$$

The MMRE calculates the mean for the sum of the MRE of n projects. Specifically, it is used to evaluate the prediction performance of an estimation model.

- **Prediction Level (PRED)**

$$PRED(L) = \frac{k}{n}$$

where L is the maximum MRE of a selected range, n is the total number of projects, and k is **number of projects** in a set of n projects whose MRE <= L. PRED calculates the ratio of projects' MREs that falls into the selected range (L) out of the total projects.
(e.g. n = 100, k =80, where L= MRE <= 30%: PRED(30%) = 80/100 = 80%)

2) *Dataset*

There are two major steps in transferring data from COCOMO 81 to SEER-SEM: first, information is converted from COCOMO 81 to COCOMO II and then from COCOMO II to SEER-SEM. The main guidelines are referred to (Madachy, Boehm, and Wu 2006; Reifer, Boehm, and Chulani 1999). In the method of the second step, 20 of the 34 SEER-SEM technical parameters can be directly mapped to 14 COCOMO II cost drivers and 1 scale factors, 1 COCOMO 81 cost driver, and 2 COCOMO 87 cost drivers. The remainder of the SEER-SEM parameters cannot be transferred to the COCOMO model, and as a result, they are set up as nominal in SEER-SEM. After transferring 93 COCOMO 81 project data points, the estimation performance with transferred data are evaluated with the estimation performance metrics. Table 1 presents the details of the prediction performance of COCOMO 81, COCOMO II, and SEER-SEM.

Table 1. Estimation Performance with Transferred Data.

	Cocomo 81	Cocomo II	Seer-sem
Mmre (%)	56.46	48.63	84.39
Pred(20%)	36.56	37.63	36.56
Pred(30%)	51.61	54.84	45.16
Pred(50%)	76.34	78.49	56.99
Pred(100%)	92.47	94.62	81.72
# of Outliers	22	20	39

The data transferring from COCOMO 81 to COCOMO II keeps the very close performance with little improvement when doing COCOMO II estimation with the transferred data. The transferring from COCOMO II to SEER-SEM causes the MMRE decreasing and the outliers increasing. Most of the new outliers come from the embedded projects whose MREs are lower than 50% before being transferred to SEER-SEM. The PRED is still stable and there is not a huge change. Overall, transferring from COCOMO 81 to SEER-SEM is feasible for our evaluation, especially when the actual project data in the format of SEER-SEM are difficult to obtain. We use the online calculator of the USC Center for Software Engineering to perform COCOMO 81 and

COCOMO II estimation. We do SEER-SEM effort estimation by two methods. One is performed by the SEEM-SEM tool (SEER-SEM for Software 7.3) which is offered by GAI, and the other is done manually by Microsoft Excel with the equations of SEER-SEM effort estimation model as presented in the section III.B. The SEER-SEM effort estimation model is also implemented as part of our research because it is part of our proposed model. The estimation performance by the SEER-SEM tool and Excel are very close. This is a way to make sure the algorithm of SEER-SEM effort estimation presented in this paper to be correct. We select the results done manually to avoid the impact from other parameters settings in the SEER-SEM tool.

The dataset of 6industrial project data points is from the COCOMO 87 model, which is slightly different than COCOMO 81, as the effort multipliers RUSE, VMVH (Host Volatility), and VMVT (Target Volatility) are not used in COCOMO 81. However, RUSE can be transferred to COCOMO II directly because it is one of the COCOMO II cost drivers, and VMVH and VMVT can be transferred to the SEER-SEM parameters DSVL and TSVL. The rest of COCOMO 87 cost drivers are matched to the corresponding cost drivers of COCOMO 81. Then, they are transferred to COCOMO II and SEER-SEM.

3) *Evaluation Cases*

After transferring the data, we conducted four main case studies to evaluate our model. These cases, which used different datasets from 93 projects, were utilized to perform training on the parameter values. The 93 project data points and the 6 industrial project data points were adopted for testing purposes. The original SEER-SEM parameter values are trained in each case. The learned parameter values of the four cases are different. This reason causes the prediction performance difference amongst the cases and the SEER-SEM. In order to assess the prediction performance of the neuro-fuzzy model, we compared SEER-SEM effort estimation model with our framework. Several performance metrics were used for the analysis of each case, including MRE, MMRE, and PRED. Accordingly, Table 2 presents the MMRE results from Cases 1 to 4, and Table 3 shows the MMRE results of the industrial project data points. Table 4 shows the PRED results of Cases 1, 2, and 3. The PRED results of Case 4 are presented in Table 5.In the tables presenting the analysis results, we have included a column named Change”, which is used to indicate the performance difference between SEER-SEM effort estimation model and our neuro-fuzzy model. For the MMRE, the prediction performance improves as the value becomes closer to zero; therefore, if the change for these performance metrics is a negative value, the MMRE for the neuro-fuzzy model is improved in comparison with SEER-SEM. Additionally, the PRED(L)” in Table 4 represent the prediction level of the selected range, referring to the definition presented in the section IV.A; a higher prediction level indicates a greater level of performance for PRED. For PRED, a negative value for the Change” indicates that our model shows a decreased level of performance as compared to SEER-SEM. Finally, the results for both MMRE and PRED are shown in a percentage format.

Table 2. MMRE of 93 Published Data Points.

Case ID	SEER-SEM	Validation	Change
C1	84.39	61.05	-23.35
C2	84.39	59.11	-25.28
C3	84.39	59.07	-25.32
C4-1	50.49	39.51	-10.98
C4-2	42.05	29.01	-13.04

Table 3. MMRE of Industrial Project Data Points.

Case ID	MMRE (%)		
	SEER-SEM	Industrial Average	Change
C1	37.54	35.54	-2
C2	37.54	47.57	10.03
C3	37.54	47.16	9.62
C4-1	37.54	33.20	-4.34
C4-2	37.54	30.39	-7.15

Table 4. PRED of Cases 1, 2 and 3.

	SEER-SEM	Neuro-Fuzzy Model					
PRED(L)	PRED (%)	C1		C2		C3	
		PRED (%)	Change	PRED (%)	Change	PRED (%)	Change
PRED(20%)	36.65	29.03	-7.62	15.05	-21.6	15.05	-21.6
PRED(30%)	45.16	37.63	-7.53	18.28	-26.88	18.28	-26.88
PRED(50%)	56.99	64.52	7.53	36.56	-20.43	38.71	-18.28
PRED(100%)	81.72	92.47	10.75	97.85	16.13	97.85	16.13

Case 1 (C1): Learning with project data points excluding all outliers

This case involved training the parameters of projects where the MREs are lower than or equal to 50%. There are 54 projects that meet this requirement. Since we wanted to perform learning without any impact from the outliers, the learning was done with 54 project data points, while 93 pieces of project data and the 6 industrial project data points were used for testing. When using the neuro-fuzzy model, the MMRE decreased from 84.39% to 61.05%, with an overall improvement of 23.35%. After testing data from the 93 projects, we used the 6 industrial project data points to perform testing. The results of this evaluation present the same tendency as the testing results with the 93 project data points: the MMRE of the neuro-fuzzy model is lower than the MMRE of SEER-SEM by 2%. With the neuro-fuzzy model, PRED(20%) and PRED(30%) decreased by 7.62% and 7.53% in comparison to the same values using SEER-SEM; however, PRED(50%) and PRED(100%) improved with the neuro-fuzzy model by a factor of 7.53% and 10.75% respectively, which indicates that the MRE of the neuro-fuzzy model, in comparison with that of SEER-SEM, contained more outliers that were less than 100% or 50%. Furthermore, the MMRE was significantly improved with the neuro-fuzzy model due to the increase of outliers that were less than 100%. By integrating the results from the MMRE, PRED, and the industrial project data points, this calibration demonstrates that the neuro-fuzzy model has the ability to reduce large MREs.

Case 2 (C2): Learning with all project data including all outliers

In Case 2, we used the data points from all 93 projects to calibrate the neuro-fuzzy model without removing the 39 outliers. The testing was performed with the same project dataset used in the training and with the 6 industrial project data points. In comparison to Case 1, this test attempted to ascertain the prediction performance when the learning involved the outliers as well as the effects of the outliers on the calibration. the MMRE using SEER-SEM comparison to the MMRE using SEER-SEM. Nevertheless, the industrial project data points caused the MMRE to worsen with the neuro-fuzzy model by 10.03%. The results of PRED demonstrate that PRED(20%), PRED(30%), and PRED(50%) decreased by more than 20%, while PRED(100%) increased by 16.13% with the neuro-fuzzy model. Moreover, these results also indicate that the neuro-fuzzy model is effective for improving the MREs that are greater than 100%. As a result, the MMRE in all of the datasets are improved when the neuro-fuzzy model is utilized. In Cases 1 and 2, the results of PRED and the 6 industrial project data points show that the neuro-fuzzy model causes large increases in small MREs while reducing large MREs. Hence, the decrease of large MREs leads to the overall improvement of the MMRE, thus showing the effectiveness of the neuro-fuzzy model.

Case 3 (C3): Learning with project data excluding part of outliers

After training, which included and then excluded all of the outliers, Case 3 calibrated the neuro-fuzzy model by removing the top 12 of 39 outliers where the MRE is more than 150%. In this case, 87 project data points are used to perform training, and the 93 project data points and the 6 industrial project data points are used for testing. The results of Case 3 are almost identical to the results of MMRE and PRED as demonstrated in Case 2. Specifically, for the neuro-fuzzy model, the MMRE of industrial project data points is worsened by 9.62%. Overall, as compared to Case 2, calibration excluding the top 12 outliers does not make a significant difference in the performance of the model.

Case 4 (C4): Learning with part of project data points

In the previous three cases, all data points from the 93 projects were used for testing. However, in Case 4, we used part of this dataset to calibrate the neuro-fuzzy model, and the rest of the data points, along with the 6 industrial project data points, were used for testing. The objective of this case was to determine the impact of the training dataset size on the calibration results. Table 2, Table 3, and Table 5 present the results.

Case 4 -1 (C4-1):

Learning with 75% of project data points and testing with 25% of project data points

This sub-case performed training with 75% of the 93 project data points and testing with the remaining 25% of these points. The project numbers for the training data points ranged from 24 to 93, while those for the testing points ranged from 1 to 23 and also included the 6 industrial project data points. To analyze the results, we compared the performance of SEER-SEM to that of the neuro-fuzzy model for Projects 1 to 23. In this case, the neuro-fuzzy model improved the MMRE by 10.98%. Furthermore, PRED(30%) and PRED(100%) with our model improved by 4.35% and 8.70% respectively. Finally, with the neuro-fuzzy model, the MREs of all 23 project data points were within 100%. In this case, the testing results of the industrial project data points are improved from the previous tests by 4.34%. These results demonstrate the effective performance of the neuro-fuzzy model in reducing large MREs.

- *Case 4 -2 (C4-2):*
  Learning with 50% of project data points and testing with 50% of project data points

Case 4-2 divided the 93 project data points into two subsets. The first subset included 46 project data points that are numbered from 1 to 46 and were used to perform testing. On

Table 5. Case 4 PRED Results.

PRED (L)	C4-1 PRED (%)			C4-2 PRED (%)		
	SEER-SEM	Neuro-Fuzzy Model	Change	SEER-SEM	Neuro-Fuzzy Model	Change
PRED(20%)	39.13	34.78	-4.35	50.00	43.48	-6.52
PRED(30%)	47.83	52.17	4.35	63.04	56.52	-6.52
PRED(50%)	65.22	60.87	-4.35	73.91	76.09	2.17
PRED(100%)	91.30	100	8.70	91.30	100	8.7

the other hand, the second subset contained 47 project data points, numbered from 47 to 93, which were used to train the neuro-fuzzy model. In comparison to Case 4-1, this test contains fewer training data points and more testing data points. Accordingly, we analyzed the performance results of the 46 project data points as estimated by both SEER-SEM and the neuro-fuzzy model. In this case, the MMRE improved by 13.04% when using the neuro-fuzzy model. Specifically, the results of PRED showed improvement from those in Case 4-1; not only were the MREs of all 46 project data points within 100%, but the MREs of most project data points were also less than 50%. Furthermore, in the testing that involved the 6 industrial project data points, the results were better than those in Case 4-1. Using the neuro-fuzzy approach, the MMRE of the 6 industrial project data points improved by 7.15%, which was the greatest improvement among all of the cases in this study.

4) EVALUATION SUMMARY

In this section, we summarize the evaluation results by comparing the analysis of all of the cases as presented in the previous sections. Fig. 6 shows the validation summary for the mmre across all of the cases. Specifically, the mmre improves in all of the cases, with the greatest improvement being over 25%.

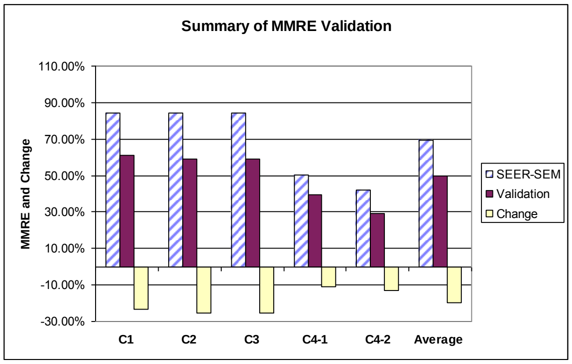


Fig.6. Summary of MMRE Validation.

Table 6 illustrates the PRED averages for SEER-SEM in all of the cases, and Fig. 7 shows the PRED averages for all of the cases using the neuro-fuzzy model. Compared to the PREDs from SEER-SEM, the averages of PRED(20%), PRED(30%), and PRED(50%) with the neuro-fuzzy model do not show improvement. However, the average of PRED(100%) is increased by 12.14%, which indicates that the neuro-fuzzy model improves the performance of the MMRE by reducing the large MREs.

Table 5. Summary of PRED Average.

	SEER-SEM	Average of Validation	Change
PRED(20%)	39.76%	27.48%	-12.28%
PRED(30%)	49.27%	36.46%	-12.81%
PRED(50%)	62.02%	55.35%	-6.67%
PRED(100%)	85.55%	97.69%	12.14%

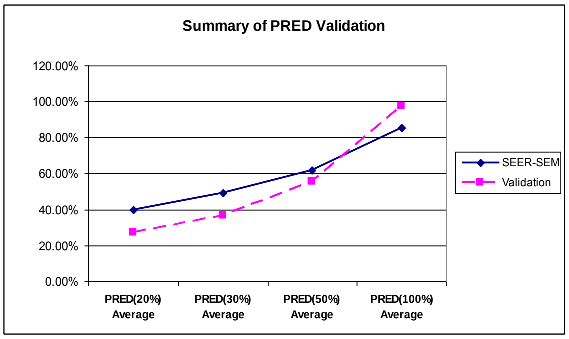


Fig.7. Summary of PRED Validation

Fig. 8 presents the MMREs of industrial project data points from all of the cases. The MMRE from Cases 1 and 4 demonstrate an improvement of no more than 7.15%. The calibrations with the outliers in Cases 2 and 3 lower the prediction performance of these two cases. Thus, for the neuro-fuzzy model, the improvement of the MMRE of industrial projects is minimal.

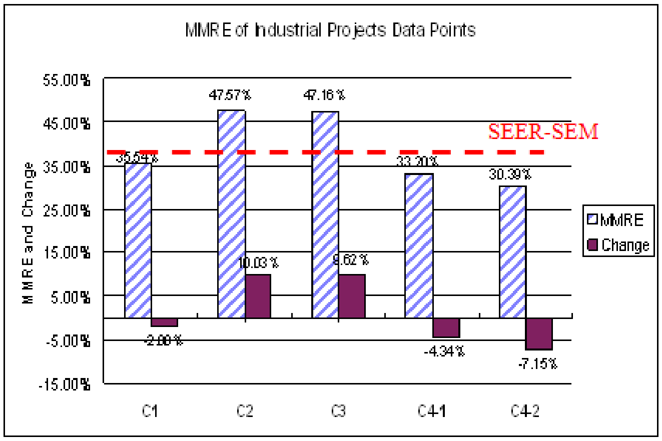


Fig.6. MMRE of Industrial Project Data Points.

V. Conclusion

Overall, our research demonstrates that combining the neuro-fuzzy model with the SEER-SEM effort estimation model produces unique characteristics and performance improvements. Effort estimation using this framework is a good reference for the other popular estimation algorithmic models. The neuro-fuzzy features of the model provide our neuro-fuzzy SEER-SEM model with the advantages of strong adaptability with the capability of learning, less sensitivity for imprecise and uncertain inputs, easy to be understood and implemented, strong knowledge integration, and high transparency.

Four main contributions are provided by this study:

a) ANFIS is a popular neuro-fuzzy system with the advantages of neural network and fuzzy logic techniques, especially the ability of learning. The proposed neuro-fuzzy model can successfully manage the nonlinear and complex relationship between the inputs and outputs and it is able to handle input uncertainty from the data.

b) The involvement of fuzzy logic techniques improves the knowledge integration of our proposed model. Fuzzy logic has the ability to map linguistic terms to variables. Accordingly, the inputs of our model are not limited to linguistic terms and can also work with numerical values. The defined fuzzy rules are an effective method for obtaining the experts‘ understanding and experience to produce more reasonable inputs.

c) There are two techniques introduced in this research: the triangular membership function and the monotonic constraint. Triangular Membership Functions are utilized to translate parameter values to membership values. Furthermore, monotonic constraints are used in order to preserve the given order and maintain consistency for the rating values of the SEER-SEM parameters. These techniques provide a good generalization for the proposed estimation model.

d) This research proves that the proposed neuro-fuzzy structure can be used with other algorithmic models besides the COCOMO model and presents further evidence that the general soft computing framework can work effective with various algorithmic models. The evaluation results indicate that estimation with our proposed neuro-fuzzy model containing SEER-SEM is more efficient than the estimation results that only use SEER-SEM effort estimation model. Specifically, in all four cases, the MMREs of our proposed model are improved over the ones where only SEER-SEM effort estimation model is used, and there is more than a 20% decrease as compared to SEER-SEM. According to these results, it is apparent that the neuro-fuzzy technology improves the prediction accuracy when it is combined with the SEER-SEM effort estimation model, especially when reducing the outliers of MRE >100%.

Although several studies have already attempted to improve the general soft computing framework, there is still room for future work. First, the algorithm of the SEER-SEM effort estimation model is more complex than that of the COCOMO model. Prior research that combines neuro-fuzzy techniques with the COCOMO model demonstrates greater improvements in the prediction performance. Hence, the proposed general soft computing framework should be evaluated with other complex algorithms. Secondly, the datasets in our research are not from the original projects whose estimations are performed by SEER-SEM. When the SEER-SEM estimation datasets are available, more cases can be completed effectively for evaluating the performance of the neuro-fuzzy model.

VI. References

1) Abran, A. and Robillard, P. N. (1996) Function Points Analysis: An Empirical Study of Its Measurement Processes. Journal of Systems and Software, Vol. 22, Issue 12: 895–910

2) Albrecht, A. J. (1979) Measuring Application Development Productivity. Proceedings of the Joint SHARE, GUIDE, and IBM Application Development Symposium: 83–92

3) Boehm, B. W. (1981) Software Engineering Economics. Prentice Hall, Englewood Cliffs, NJ

4) Boehm, B. W., Abts, C., Brown, A. W., Chulani, S., Clark, B. K., Horowitz, E., Madachy, R., Reifer, D., and Steece, B. (2000) Software Cost Estimation with COCOMO II. Prentice Hall, Upper Saddle River, NJ

5) Boehm, B. W., Abts, C., and Chulani, S. (2000) Software Development Cost Estimation Approaches – A Survey. Annuals of Software Engineering: 177–205

6) Chulani, S. (1999) Bayesian Analysis of Software Cost and Quality Models. Dissertation, University of South California

7) Damiani, E., Jain, L. C., and Madravio, M. (2004) Soft Computing in Software Engineering. Springer, New York, NY

8) Fischman, L., McRitchie, K., and Galorath, D. D. (2005) Inside SEER-SEM. Cross Talk – The Journal of Defense Software Engineering: 26–28.

9) Galorath, D. D. and Evans, M. W. (2006) Software Sizing, Estimation and Risk Management. Auerbach Publications, Boca Raton, NY

10) Galorath Incorporated (2001) SEER-SEM User‗s Manual

11) Galorath Incorporated (2005) SEER-SEM Software Estimation, Planning and Project Control Training Manual

12) Gray, A. R. and MacDonell, S.G. (1997) A Comparison of Techniques for Developing Predictive Models of Software Metrics. Information and Software Technology, Vol.39, Issue 6:425–437

13) Ho, D. (1996) Experience Report on COCOMO and the Costar Tool from Nortel‗s Toronto Laboratory. the 11th International Forum on COCOMO and Software Cost Modeling

14) Hodgkinson, A. C. and Garratt, P. W. (1999) A NeuroFuzzy Cost Estimator. Proc. 3rd Int Conf Software Engineering and Applications (SAE): 401–406

15) Huang, X. (2003) A Neuro-Fuzzy Model for Software Cost Estimation. Dissertation, University of Western Ontario

16) Huang, X., Ho, D., Ren, J., and Capretz, L. F. (2005) A Soft Computing Framework for Software Effort Estimation. Soft Computing: 170–177

17) Huang, X., Ho, D., Ren, J., and Capretz, L. F. (2006) Improving the COCOMO Model Using A Neuro-Fuzzy Approach. Applied Soft Computing: 29–40

18) Huang, X., Ho, D., Ren, J., and Capretz, L. F. (2008) System and Method for Software Estimation. USA Patent No. US-7328202-B2

19) Jang, J. R., Sun, C. and Mizutani, E. (1997) Neuro-Fuzzy and Soft-Computing. Prentice Hall, Upper Saddle River, NJ

20) Jensen, R., Putnam, L., and Roetzheim, W. (2006) Software Estimation Models: Three Viewpoints. Software Engineering Technology: 23–29

21) Jones, T. C. (1998) Estimating Software Costs. McGraw Hill, Hightstown, NJ

22) Madachy, R., Boehm, B., and Wu, D. (2006) Comparison and Assessment of Cost Models for NASA Flight Projects. 21st International Forum on COCOMO and Software Cost Modeling

23) Nauck, D., Klawonn, F., and Kruse, R. (1997) Foundations of Neuro-Fuzzy Systems. John Wiley & Sons, Inc., New York, NY

24) Nguyen, H. T., Prasad, N. R., Walker, C. L., and Walker, E. A. (2003) A First Course in Fuzzy and Neural Control, Chapman& Hall /CRC, Boca Raton, FL

25) Panlilio-Yap, N. and Ho, D. (2000) Deploying Software Estimation Technology and Tools: the IBM SWS Toronto Lab Experience. the 9th International Forum on COCOMO and Software Cost Modeling

26) Putnam, L. H. and Myers, W. (1992) Measures for Excellence. Prentice Hall, Englewood Cliffs, NJ

27) Reifer, D. J., Boehm, B. W., and Chulani, S. (1999) The Rosetta Stone – Making COCOMO 81 Estimations Work with COCOMO II. CrossTalk The Journal of Defence Software Engineering: 11–15

28) Takagi, T. and Sugeno, M. (1986) Derivation of Fuzzy Control Rules from Human Operator‗s Control Action. Proc. of the IFAC Symp. on Fuzzy Inf. Knowledge Representation and Decision Analysis: 55 – 60

29) USC Center for Software Engineering (2006) Cost Model Comparison Report. Dissertation, University of South California

30) Wong, J., Ho, D., and Capretz, L. F. (2008) Calibrating Functional Point Backfiring Conversion Ratios Using Neuro-Fuzzy Technique. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 16, No. 6: 847 – 862

31) Xia, W., Capretz, L. F., Ho, D., and Ahmed, F. (2008) A New Calibration for Function Point Complexity Weights. International and Software Technology, Vol. 50, Issue 7-8: 670–683

Centralized Access Distribution Security System a Solution to Data Breach Security Problem

Syed Ehsen Mustafa[1], Irfan Anjum Manarvi[2]

GJCST Classification
H.2.7, C.2.0

Abstract*- The focus of this paper is to identify critical data security problems after analyzing the data breach incidents reported during the years 2006 to 2009, in order to provide an effective and efficient solution by proposing a security system that would provide protection against data beaches. In this Paper the analysis of databases for security breach incidents provides a good review for type of businesses that are affected by the data breaches, the type of data targeted for data breach attacks, frequency for type of breaches and attacks that are used to compromise the data security. After the identification of higher frequency threats and ways in which data is compromised, a solution has been provided using the problem solving techniques. The proposed Centralized Access Distribution (CAD) Security System is an efficient and effective solution based on the nine components essential to provide a solution to the identified problems of data breaches. CAD focuses on providing a configurable security system that would provide data confidentiality, data integrity, intrusion detection and prevention, automation of security procedures while monitoring the states of objects and subjects available in access control lists and features of data logging for threat analysis, thus providing a complete solution to data breaches on the rise.

Keywords- Data breach, Information security, Centralized Access distribution, RBAC model, Network Security.

I. Introduction

This paper attempts to provide a solution for the problems identified after the detailed analysis of the records (i.e. Data breach incidents reported) available at [12]. In this paper the analysis has been done on the records for year 2006 to 2009 in order to identify the data breach problems being faced by the different business types including Private Businesses, Government Organizations, Educational Institutions and Medical Institutions.In order to solve the problems of security breach and data breach most of the security systems are based on Access control models. Two commonly known access control models are Discretionary Access control (DAC) and Mandatory Access Control (MAC). DAC [1, 2, 4, 13] provides basic security on information flows and is useful in development of commercial security systems, but the problem with DAC is that it is vulnerable to the attacks of unauthorized access, for example it does not provide protection against attacks like Trojans. In comparison to DAC, MAC [1, 2, 4] does provide the functionality of unauthorized access, for example it does not provide protection against attacks like Trojans. In comparison to DAC, MAC [1, 2, 4, 14] does provide the functionality of preventing unauthorized user access, but the limitation of MAC is that due to the level of security achieved by MAC and the complexity of its implementation, it is mainly used for military security systems.Role Based Access Control Model (RBAC) was proposed in early 1990‘s by American National Standards and Technology Research Institute. RBAC model [1, 2, 3, 4, 5, 15] has been powerful in controlling information flows as compared to the functionalities provided by the traditional DAC and MAC models. RBAC model has proved to be a generic model that could be extended to develop new security models and security systems where the data security criteria of data confidentiality, data integrity, intrusion detection and prevention are fulfilled.In this paper study of different extensions of RBAC model have been done, where each extension solves a problem of data security for different working environments. The solution provided in this paper as a Centralized Access Distribution is based on the combination of features presented in different RBAC model extensions discussed below.In [1] a feature of task based permission management has been added to existing RBAC model. This feature is helpful in managing the access control of users on multiple devices on the network. Addition of this feature ensures the authentication and authorization of users on the network at abstraction layer and controlling dependencies of subjects and objects involved in the network management system.Adding the functionality of automated prevention and monitoring to existing RBAC model by using State Transfer Based Dynamic policy would result in development of an effective and efficient access control system [2]. The concept of this policy is to assign access priorities to systems on network and monitor the states of the system that are active on the network. The state of the system can change dynamically based on the policy defined for the active systems. The state based transfer controls the unauthorized requests of active systems and grant access to system of high priority, thus integrating user authentication and access control to achieve better security [16]. Combination of object oriented approach with RBAC model results in controlling information flows and providing intrusion prevention functions [4]. The concept of Access control List (ACL) is used to control information leakage and unauthorized access to databases of data under protection.As Business environments are targeted for compromising the data confidentiality and information leakage, [6] provides concept of combined Network security and data security that

About[1]- Syed Ehsen Mustafa is MSc Engineering Management Student at CASE, Centre of Advanced Studies in Engineering, Islamabad Pakistan ehsen67@yahoo.com.
About[2]- Irfan Anjum Manarvi is Associate Professor at Iqra University Islamabad Campus, Islamabad Pakistan. irfanmanarvi@yahoo.com.

would be effective in securing the business environments. The key is to group the data and processes (that are available in a Business information networks) in to authorized access level Sets. Once the access levels are defined a function would be implemented that would assign the access rights to each group according to their access level [17]. Such a controlled implementation would control unauthorized access within the network and prevent any external unauthorized access. Implementing ACL in security systems could help in improving Network security [18]. An access control approach of using encryption techniques can be used to prevent an unauthorized access of data [7, 8]. The policy of hiding data from unauthorized user such that only legitimate users can see and access the information can be used as a feature addition to the existing RBAC model for enhanced security. This feature ensures secure information exchange between different processes running on a network.Information leakage is a major threat for any social networking environment. The solution provided in [9] not only considers the logical access of the system but also the physical access in order to ensure security of data on network. [19] Introduces the feature of automated monitoring for detecting intrusions by controlling the states of all components on the network under protection.For the development of a security system it is essential to follow a formal security model, where a model can be extended to achieve desired level of security [10].It is also important that a security system should be configurable, in order to update the security processes against newly identified threats [11], just like the Preventive Information security management system providing strong intrusion prevention capabilities. But in addition to intrusion prevention features [20] discusses about the control of authorized user in order to avoid insider malicious activities, as a network without insider security is still vulnerable even if it is protected for outside security threats.In this paper Section II describes the methodology used for data analysis for identification of problems and the techniques used for solving the problems identified. Sections III to XIII are about the analysis of data breach incidents. Section XIV explains CAD system in detail. Section XV is presents the summary of all the finding of analysis and finally section XVI discusses conclusions.

II. Methodology

In this paper a detailed analysis of security breach incidents has been done. In order to analyze the real time data breach incidents a database of data breach records was downloaded from [12] for a period of four years 2006 to 2009. The data acquired from this website was used to analyze the data breach incidents. The variables used for analysis are four business domains targeted for data breach attacks, eleven breach types used for attacks, three different data types that were targeted for data breach attacks, total number of people affected by these attacks, four different sources of attacks and variable for data recovery after data breach attacks. In general tools used for analysis were bar plots, time series plot, doughnut plots, pie charts and pivot tables.First it was analyzed that how many people were affected due to data breach incidents during years 2006 to 2009. A pie chart was drawn showing the yearly percentages of people getting affected by data breach attacks, where the year 2009 was observed to be the year where maximum people were affected in comparison to other years.Then a variable of business types was analyzed by drawing a pie chart of percentages showing that private businesses are the most affected by these incidents. A pivot table of business domains versus total affected was created in order to analyze the frequency for number of people getting affected in each business domain. A bar plot was also drawn to graphically observe the relation of total affected versus business domains.After identifying the variable of business domains, the variable of breach types was analyzed by creating a pivot table and a bar plot showing the frequency of breach type being used most of the time to launch data breach attacks. The fact that the total number of people getting affected by each breach method was also analyzed by drawing a pie chat of percentages for total affected versus breach type.After analyzing the most critical breach types, analysis of different data types was done in order to evaluate how different breach methods affect different data types and to identify the data type being targeted maximum number of time during these years. For this purpose pivot tables and bar plots were drawn, with a time series plot also drawn for analyzing the trend of data types being affected during years 2006 to 2009 such that the facts gathered will be helpful while developing a new security systems. Comparison of data type versus business type was also done by drawing pivot table and bar plots to analyze the type of data getting affected corresponding to each business domain mentioned in records. Next the analysis of total affected versus data types was done in order to analyze that how many are affected with respect to each data type mentioned in records.The variable of Attack types was also analyzed by in order to identify the sources of these data breach incidents. Again this variable was analyzed by using pivot tables and bar plots. Time series plot was also drawn in order to identify the trend of attacks being launched during last four years. Finally the variable of data recovery was also analyzed by drawing a doughnut plot showing the percentages of data recovered and the data that was unable to get recovered.After analyzing the and identifying the high frequency threats, a solution was proposed that would result in an security system providing solution to all identified high frequency threats.In order to propose a solution study of different security models was done. As the goal was to present a security system, it very important to select a traditional model on which the proposed security system should be developed. So the studied security models were then evaluated on the criteria identified after analysis of data breach incidents. The different alternatives were first rated using Thomas Saaty‘s Matrix and then selection was done by using SFF matrix.

III. Analysis of total affected by data breach incidents

It is important to analyze that the total number of people affected by the data breach incidents during years 2006 to 2009 in order to evaluate the depth of this problem that is affecting millions. Following Figure-1 shows the percentages of total number of people affected by data breach incidents every year.

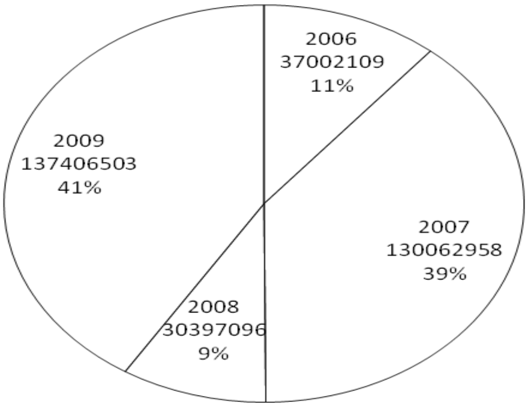


Fig.1. Percentages of Total affected by Data Breaches

Inferences drawn from the above figure are as follows:

(1) A total of 334868666 have been affected by the data breach incidents during years 2006 to 2009.

(2) About 41% got affected during year 2009 being the highest number of people affected in comparison to years 2006 to 2008.

(3) It shows that there is an increase in number of people getting affected by data reach incidents in 2009

Next it‘s required to analyze domains/businesses that have been affected, the methods that been use used to affect 334868666 people and the type of data that has been targeted for data breach.

IV. Analysis of different Business types affected by data breach incidents

The variable Business Types in the records represents four different types of business domains that are affected by the data breach security problems. The data breach incidents are reported for the following business domains:-

1) Private businesses (Biz), that includes corporate offices, IT firms, Banks, Leisure and Food Industry.
2) Educational Institutions (Edu).
3) Government Organizations (Gov).
4) Medical Institutions (Med).

Following Figure-2 shows a pie chart for the percentages of above mentioned business domains affected by the data breach incidents during years 2006 to 2009.

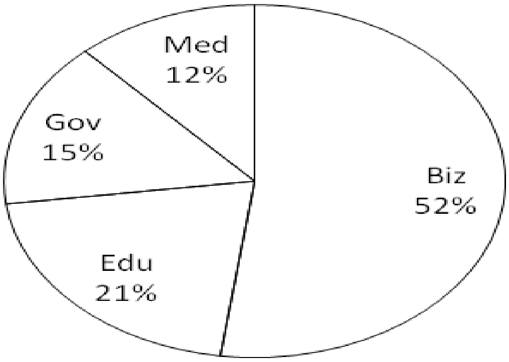


Fig.2. Percentages of Types of Business domains affected

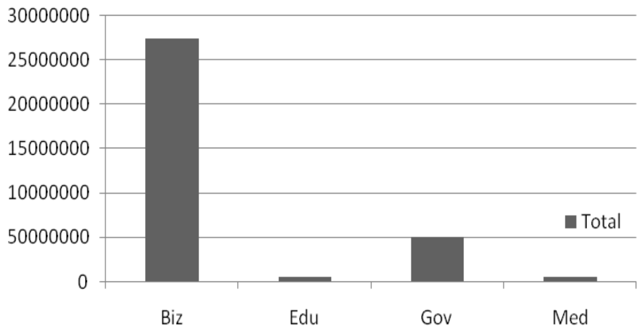


Fig.3. Total Affected Vs Business Types

Business Types	Sum of Total Affected
Biz	274500860
Edu	5287712
Gov	50513093
Med	4567001
Grand Total	334868666

Table 6.Pivot Table for Sum Total Affected Vs Business Types

Inferences drawn from the above figures are as follows:

1) Private businesses are the most affected by data breach incidents with 52% as the highest percentage of incidents being reported from this domain.
2) Second highest domain affected by the data breach incidents is Educational institutions with a percentage of 21%.
3) According to Table-1 274500860 affected are related to Private business domains. It shows that Private business domains are more vulnerable to data breach attacks as compared to other business domains.
4) As inferred that Private businesses are the most affected can be verified by observing the histogram in Figure-3 and Table-1 that significant amount of people affected are from Private business domains.

As private businesses are the most affected type, it‘s required to analyze different breach methods used to attack different business domains.

V. ANALYSIS OF BREACH TYPES REPORTED IN RECORDS

The variable Breach Type represents the methods used for attempting data breach in different business domains. Following are different breach types reported:-

(1) Data breach using web based attacks.
(2) Data breach based on fraud or scam (usually insider-related), social engineering.
(3) Data breach using hacking techniques. Computer-based intrusion including data that should not be exposed publically.
(4) Data breach because of exposure to personal information via virus or Trojan (i.e. keystroke logger, possibly classified as hacking).
(5) Data breach by snail mail. Scenario can be of personal information in "snail mail" getting exposed to unintended third party.
(6) Data breach because of disposal document i.e. information disclosure because of documents not being disposed of properly.
(7) Data breach because of information disclosure due to lost or stolen document.
(8) Data breach via stolen laptops or stolen computers.

Table-1 is a pivot table showing the sum of different breach types during years 2006 to 2009, Figure-2 is the depiction of Table-1 in terms of highest numbers of method used to compromise the data.

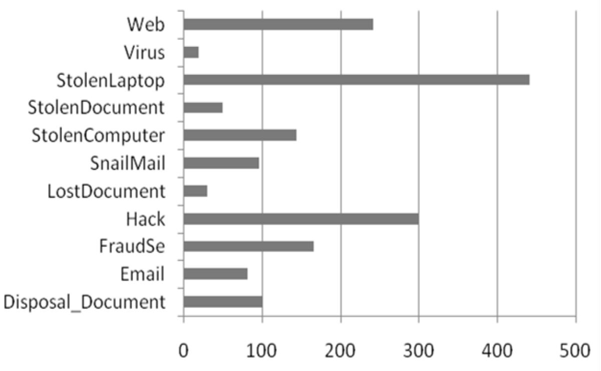


Fig.4. Bar plot showing sum of all Breach Types

Row Labels	Count of Breach Type	Sum of Total Affected
Disposal_Document	101	454740
Email	81	166279
FraudSe	165	23827397
Hack	299	254990453
LostDocument	30	1143278
SnailMail	95	7808681
StolenComputer	143	31292269
StolenDocument	49	351954
StolenLaptop	441	12420046
Virus	18	56230
Web	241	2357339
Grand Total	1663	334868666

Table 7.Pivot Table for Sum of Breach Types

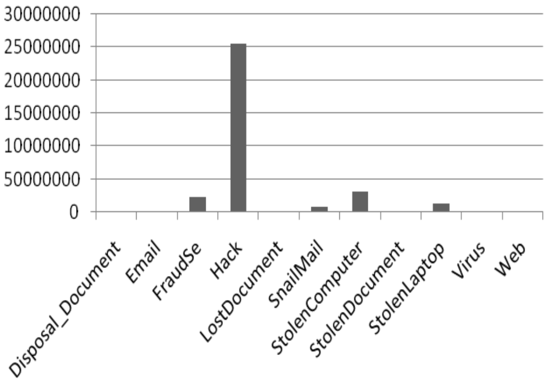


Fig.5. Frequency of Total Affected Vs Breach Types

Inferences drawn from the above pivot table and figures are as follows:

1) Data breach due to stolen laptops has the highest count of 441, showing that most of the times a major reason of data being compromised is due to storage of confidential data on laptops.
2) Second highest methods used for data breach are hacking and web based attacks as the reported incidents for hacking are 299 and web based attacks are 241.
3) Viewing the bar plot we can see that most of the data is compromised or leaked because of unprotected data available on stolen laptops, computers, documents or lost documents.
4) According to the pivot table, method of Hacking has affected about 255499053 as the highest number in comparison to other methods. Histogram in Figure-5 is the depiction of pivot table-2 for total affected versus breach types and it verifies the inference that data breach due method of hacking has affected the significant amount of people.
5) Observation shows that the lack of features for protected data availability, and data confidentiality in a security system could lead to data breach incidents on stolen laptops and computers. A security system must ensure that the data being used by the authenticated employees must be protected and remain confidential after their use.
6) Hacking as observed to be the second most method used for data breaches but the number of people it has affected is greater than any other method discussed in this analysis. This shows that a security system should have control over malicious and ambiguous events occurring within the network.
7) After analyzing different methods used for data breaches, different data types that are vulnerable to data breach needs to be analyzed in order to

classify data types targeted using the methods discussed in above inferences.

VI. Analysis of Data types Affected

The variable Data Types represents different types of data that are under attack of data breaches. Following is the description of different types for data, reported to have been compromised:-

1) Financial Information (FIN) including Credit Card numbers, Bank Account information etc.
2) Data related to Social Security Numbers (SSN).
3) Medical data (MED) including patient history, employee Medical History available in HR records etc.).
4) Mixed data including both financial data and social security numbers.
5) Mixed data including both financial and medical data.

Following Figure shows the sum of different data types being targeted for data breach attacks.

	Data Types					
Years	FIN Data	MED Data	MED and FIN Data	SSN and FIN Data	SSN Data	Grand Total
2006	60	20	2	51	247	380
2007	67	22	1	59	215	364
2008	116	48	5	72	296	537
2009	98	58	5	65	156	382
Grand Total	341	148	13	247	914	1663

Table 8.Pivot Table for Data Types

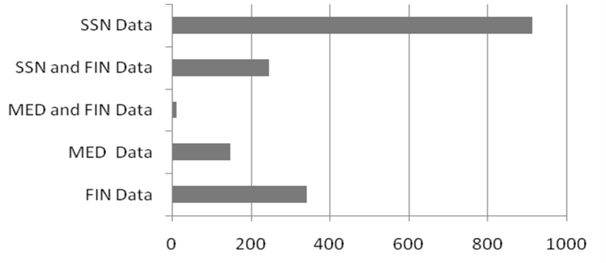


Fig.6. Bar Plot showing sum of all Data Types

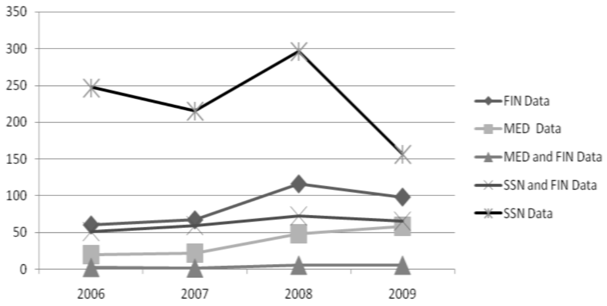


Fig.7. Time Series plot for Data Types

Inferences drawn from the above pivot table and figures are as follows:

1) Figure-6 shows different frequencies of data types presenting the fact that during years 2006 to 2009 most of the data breach attacks targeted Social Security Numbers data.
2) Out of 1663 incidents reported 914 were related to Social Security Numbers, which is about 55% in total.
3) Figure 7 representing a time series plot for different data types also shows that every year from 2006 to 2009 attacks on Social Security Numbers were the high frequency reported incidents.
4) According to Figure 7 and pivot table 3 highest numbers of incidents were recorded in year 2008 and after analyzing Pivot Table-3 it can be observed that total numbers of incidents dropped by 28% in year 2009. But still incidents related to Social Security Numbers were the highest according to the observations in Figure 7.

VII. Comparison of Business type Vs Data type

Table-4 shows the comparison of two variables from the data breach incidents recorded.Figure-8 is the depiction of pivot table-4.

	Data Types					
Business Domains	FIN Data	MED Data	MED and FIN Data	SSN and FIN Data	SSN Data	Grand Total
Biz	295	21	8	199	345	868
Edu	10	24		15	296	345
Gov	25	16	1	24	187	253
Med	11	87	4	9	86	197
Grand Total	341	148	13	247	914	1663

Table 4.Sum of Data Types Vs Business Domains

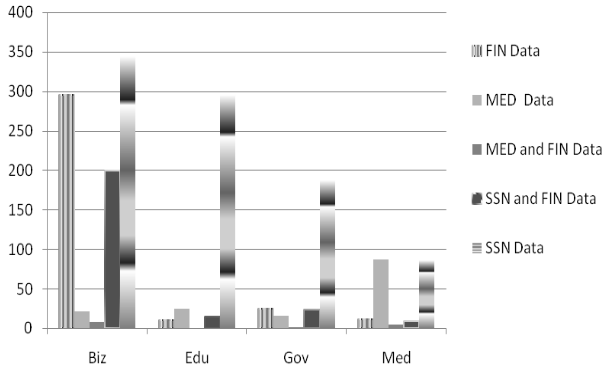


Fig.8. Comparison of Business Type vs. Data Type

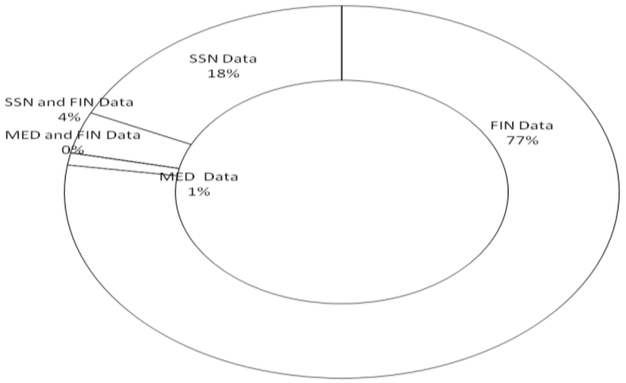


Fig.9. Percentages of Total Affected Vs Data Type

Inferences drawn from the above pivot table and figure are as follows:

1) Figure-8 shows that data breach attacks on Private businesses targeted Social Security Numbers data and financial data with recorded 345 and 295 frequencies of incidents respectively. These records are the highest in comparison to other businesses.
2) In Educational Institutions and Government Organizations frequency of attacks on Social Security Numbers data is recorded as the highest in number (i.e. 296 and 197).
3) According to the above inferences information leakages of Social Security Numbers data is on the rise during years 2006 to 2009.It shows that the features of data hiding and data confidentiality should be supported by security systems in order to protect this large amount of data.
4) Although a large number of incidents reported were related to attacks on Social Security Numbers data but according to Figure-9 doughnut plot most of the people are affected by attacks on financial data. Doughnut plot in Figure-9 shows that 77% of 334 million people are affected by attacks on financial data.

I. Comparison of Breach Type Vs Data Type

Figure-10 represents the frequencies of data type against the two highly recorded breach types in order to analyze how different breach types target different types of data.

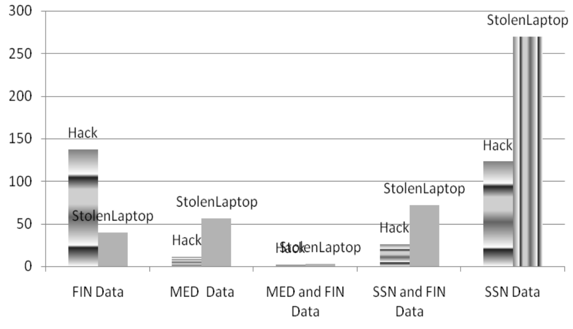


Fig.10. Comparison of Breach Type Vs Data Type

Inferences drawn from the above pivot table and figure are as follows:

(4) Figure-10 shows that data breach from stolen laptops and hacking affects Social Security Numbers databases, where SSN data breach incidents due to stolen laptops is above 250 and due to hacking are above 100.
(5) Incidents where financial data type is compromised are below 50 in case of stolen laptops and above 140 in case of hacking as a method of data breach attack.
(6) The method of hacking and information leakage due to stolen laptops has also affected medical related data but the amounts of incidents recorded are less in comparison to other data types.

In the next section analysis of attack type needs to be done in order to identify the security measures that are needed to be implemented for prevention of data breach attacks.

II. Analysis of Attack Types

The analysis of variable Attack type will help to identify the security procedures that need to be taken to prevent the data breach attacks. Following is the description of different categories of attacks as per records:-

1) Outside attacks (hacking, malware, viruses)
2) Inside Malicious attacks (Intruder attacks, Fraud, scams).
3) Inside Accidental attacks (Due to untrained employees, careless management of data, documents).

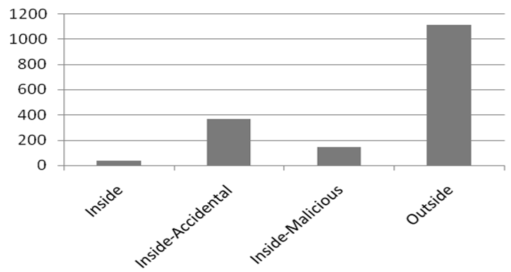


Fig.11. Frequency of Attacks Types

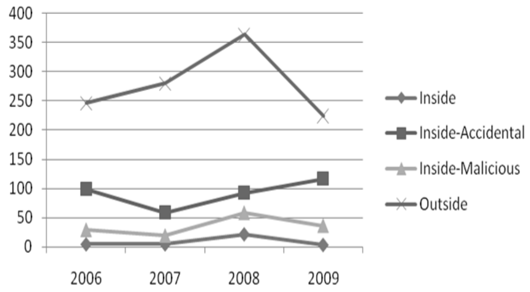


Fig.12. Time series plot for Type of attacks 2006 to 2009

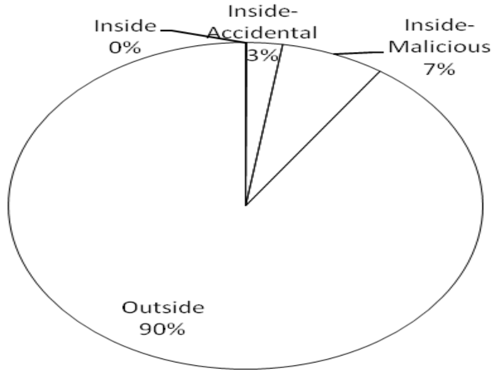


Fig.13. Percentages of Total Affected Vs Attack Type

Inferences drawn from the above pivot table and figure are as follows:

1) Figure-11 shows that maximum of data breach attacks are from the category of outside attacks, with a frequency of 1113 incidents recorded as outside attacks. This shows that Automation of monitoring and data breach prevention techniques needs to be present in a security system in order to take care of outsider attack threats.
2) Second highest number (i.e. 368 out of 1663) of incidents recorded are related to insider malicious attacks category, which shows that accidental exposure of confidential data is also a case of data breach. Such a problem requires automated monitoring to prevent accidents and development of security training policies.
3) Figure-12 shows the time series plot verifying that outsider attacks are highest during years 2006 to 2009 where Figure-13 shows that outsider attacks has affected 90% of 334 million people. So outsider attacks are a major problem for data security and have been affecting millions of people during years 2006 to 2009.
4) As inferred that from Figure-12 that frequency of insider accidental attacks is the highest but according to Figure-13 pie chart insider malicious attacks (7%) has affected 4% more people then insider accidental attacks (3%).So it more critical to solve problem of insider malicious attacks then insider accidental attacks.

X. Comparison of Attack Type Vs Data Type

Following comparison helps in analyzing how different attack types have affected different data types during years 2006 to 2009.

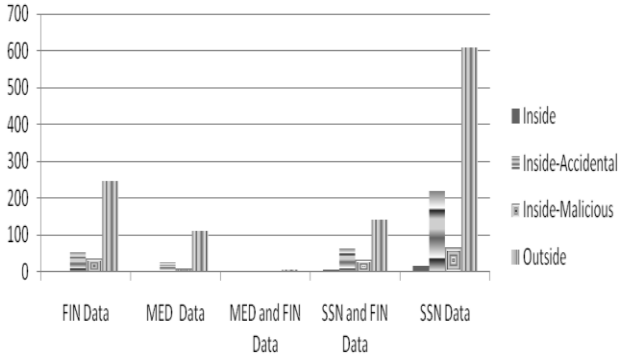


Fig.14. Comparison of Attack Type Vs Data Type

Inferences drawn from the above pivot table and figure are as follows:

1) As from the previous inferences we already know that SSN data is the most targeted data for data breach attacks, observation on Table-6 shows that 600 plus records of incidents as the highest in number are related to the category of outside attacks used to compromise SSN data.
2) Figure-14 shows that inside accidental attacks and outside attacks both are a major source of attacks used to compromise SSN data.
3) By observation inside accidental attacks and outside attacks again are the major source attack types for financial data recorded as the second highest data type under data breach attacks.

XI. Comparison of Attack Type Vs Business Type

Following comparison helps in analyzing the high frequency breach methods used to perform data breach attacks.

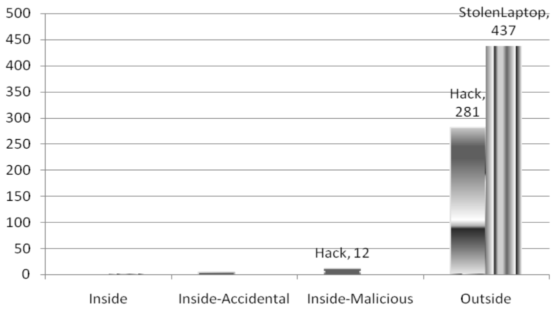


Fig.15. Comparison of Attack Type Vs Business Type

Inferences drawn from the above figure are as follows:

1) Figure-15 shows that outside attacks have affected all four types of business domains with a maximum amount of incidents recorded as outside attacks, where above 590 incidents as the maximum number of attacks have been reported for private businesses.
2) Figure-15 also verifies the previous inferences that the category of inside accidental attacks as the

second highest in number has also affected all four business domains.

XII. Comparison of Attack Type Vs Breach Type

Following comparison helps in analyzing the high frequency breach methods used to perform data breach attacks.

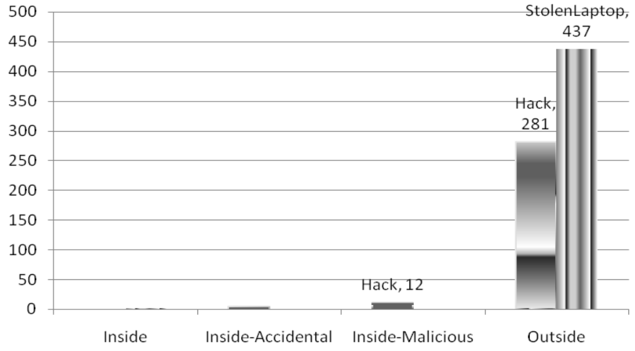


Fig.16. Frequency of Attack Type Vs Breach Type

Inferences drawn from the above figure are as follows:

1) Figure-16 shows that outside attacks with maximum number of 437 incidents recorded used stolen laptop breach method (i.e. previously inferred to be the most commonly used breach method).
2) Figure-16 shows that outside attacks with frequency of 281 out of 1663 records used hacking as a breach method (i.e. previously inferred to have affected maximum number of 25 million people).

a) Analysis of Data Recovered

The variable of data recovered shows the statistics for data recovery out of all the incidents reported. The facts drawn from the following figure will help in identifying the need of integrating data backup and data logging features in a security system.

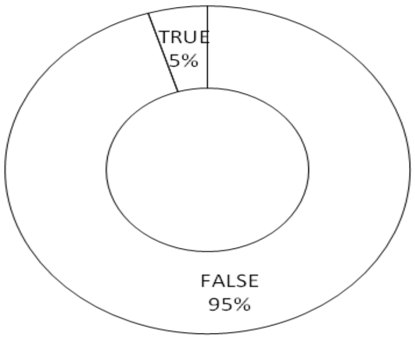


Fig.17. Doughnut plot showing percentages of Data Recovered

Inferences drawn from the above figure are as follows:

1) Figure-17 doughnut plot showing the percentages for data recovered in reported incidents where in 95% of the recoded data breach incidents data was not recovered which shows that the features of data logging and feedback are missing from most security systems.
2) It means that this 95% of data that was lost shows the scale of information leakage problem that is related to 334 million people (according to the above discussed statistics) needs one generic solution against the threats of hacking, network intrusion, and accidental information exposure.

XIV. Centralized Access Distribution System

The detailed analysis of data breach incidents in the above discussions shows that in current business environments following critical problems are observed.

1) Problem of data confidentiality (mostly observed in private business domains).
2) Controlling unauthorized access in to the networks (as maximum numbers of incidents recorded are outside attacks i.e. outside network intrusion).
3) Prevention of breach methods like hacking.
4) Data logging and feedback mechanisms.
5) Training policies to avoid accidental exposure of confidential data.

The goal of this newly proposed Centralized Access Distribution (CAD) system is to provide solution for the above identified problems. In order to develop a security system it is essential that a security system should be developed based on some traditional authorized security model where basic elements of providing security are available. The importance of security model compliance is that it provides a structured base one can use and enhance for achieving desired security.So for the proposal of CAD security system, first a traditional security model should be selected. As discussed in the Introduction section there are three commonly known traditional data security models being used for development of security systems and can be extended for proposal of new security model to be used for specific environments.In this paper three models that are selected for evaluation are Discretionary Access model (DAC), Mandatory Access Model (MAC) and Role Based Access Control (RBAC) Model. Evaluation of these selected models is done on the defined criteria using Thomas Saaty‘s Matrix. The following criteria are the basic requirement of proposed CAD model:-

1) Does the model Provide data confidentiality features?
2) Does it provide data Integrity?
3) Does it provide protection against unauthorized intrusion?

Alternatives	DAC	MAC	RBAC	SUM	RANK
DAC		0	0	0	3rd
MAC	1		0	1	2nd
RBAC	1	1		2	1st

Table 5.Thomas Saaty's Matrix for Model evaluation

RBAC model stands at first position as it fulfills the criteria required for developing basic structure of proposed CAD model, where as DAC and MAC models do support data security but they have their own limitation as discussed and mentioned in the Introduction section of this paper.For the selection of best alternative SFF Matrix is used. Criteria for SFF are as follows:

1. *Criteria for suitability*

Does this model provide data security mechanisms for secure access control over the network and data sources?

2. *Criteria for Feasibility*

Implementation complexity of the Model (Is it easily implementable and generic to be used commercially).

3. *Criteria for Flexibility*

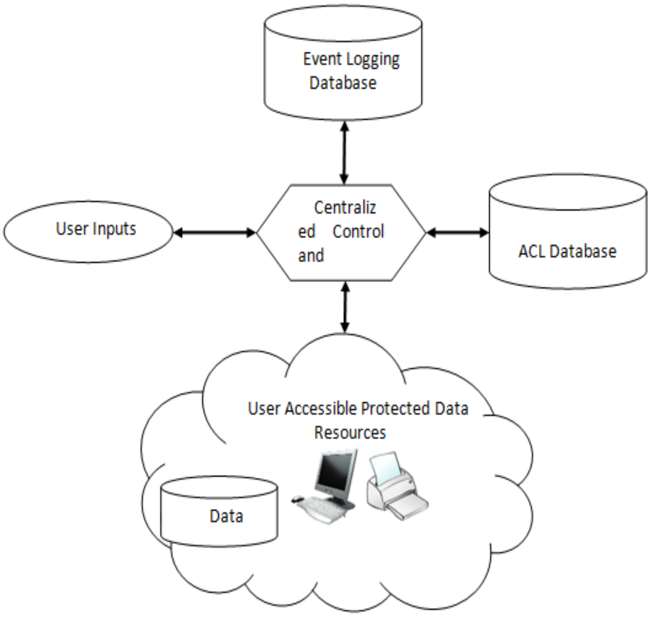
Is it easily extendable for achieving desired security (i.e. addition of new features)?Now the selected Alternatives (i.e. security models) will be awarded points from 1 – 3 with 1 being the lowest point and 3 being the highest in SFF Matrix below.

Alternatives	Suitability	Feasibility	Flexibility	Total
DAC	1	1	1	3
MAC	2	1	1	4
RBAC	2	2	2	6

Table 6. SFF Matrix for selection of Security Model

On the basis of results obtained from SFF Matrix, RBAC model has a higher total of 6. In comparison to the results of Thomas Saaty's Matrix RBAC model was ranked at first position and here at SFF Matrix also RBAC has proved to be the best alternative, so RBAC model will be used for the development of Centralized Access Distribution system.After the analysis and identification of core problems from the data breach incident database, the proposed CAD system is defined based on nine components that would serve as the features of the CAD system. Each of these nine components provides a solution to the real time industrial problems identified after analysis. Figure 18 shows the architecture of CAD security system proposed as a solution to Identified problems

Fig.18. Architecture of CAD Security Syste



CAD Security System consists of three types of Access control Lists (ACL):-

1) Access control list for Users registered into the security system with a tag representing active and inactive users.
2) Access control list for roles and permissions associated with each user to access resources.
3) Access control list for resources that are under observation and available on the network.

As shown in Figure-18 a database for access control lists is available and a list a database of event logging is also available. One centralized module functions to control the user inputs (for accessing the network resources) and monitor the activities going on the user accessible protected data resources. Whenever there is a request from the user to access a protected resource, the central monitoring module verifies user registration from user ACL and then verifies the request such that the user is authorized to access the protected resource. Once the authorization process is complete user is allowed to access the desired resource.

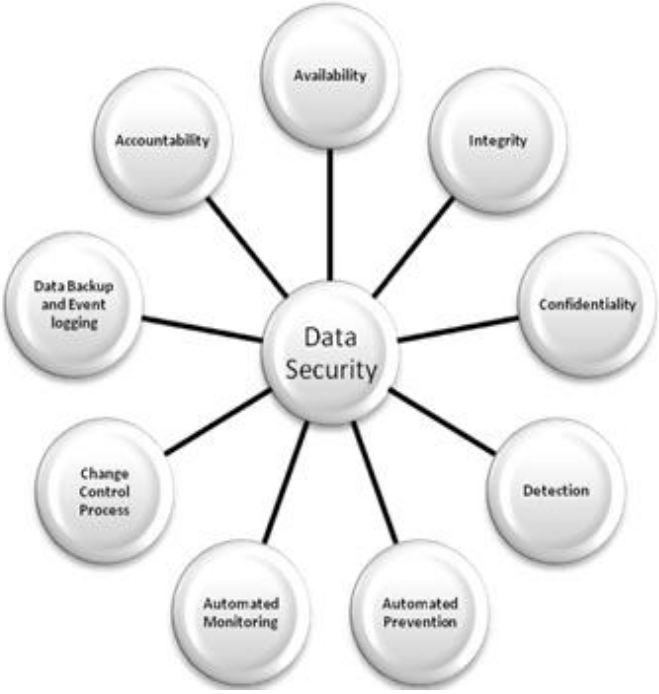


Fig.19. Components of CAD Security System

Nine different components that are used to ensure the security of a network and it resources are described below.

4. *Confidentiality*

This feature ensures the protection of data throughout organization's information architecture.As data confidentiality is the basic feature that can be ensured by the implementation of encryption techniques while transferring data on the network. This is equivalent to using data hidden policy as the data can only be decrypted and viewed by authorized user.According to the analysis done on security breach incidents database data loss due to stolen laptops was identified as high frequency breach type. The proposed CAD security system provides a solution to this problem by centralizing the data to be used by the laptop users such that the data does not needs to be present on the laptop it can be accessed from the central server using a Virtual Private Network connection from the authorized laptop. Such an arrangement of data usage is helpful in case if a laptop is stolen there will be no data available on the laptop because in current scenario laptop will only be acting as a data processing machine taking encrypted data from central server and then saving the processed data back to central server instead of saving it on laptop.

5. *Integrity*

This feature ensures unauthorized alteration or destruction of data and data providing services. The implementation of this feature refers to the use of Access control list for identification of authorized users such that the resources on the network are dedicated to users according to their access levels. Therefore CAD system based on the principles of RBAC model maintains centralized access control list of authorized users, a list of access levels assigned to each user and a list of resources assigned to each user. This structure helps in ensuring integrity of data under observation.

6. *Availability*

This feature ensures that the data and other services on the secured network are always available for authorized access.The implementation of this feature is ensured by the implementation of central access control list database that would always be available to the monitoring module of the CAD security system. The feature of data confidentiality requires data resources on a centralized data server where the CAD security system needs to ensure that the data resources are always available for processing.

7. *Accountability*

This feature ensures control over malicious and ambiguous events occurring within in a network. The implementation of this feature is ensured by assigning access control roles to the authorized user active in the network. As access control lists are maintained for each user, the monitoring module can easily detect an event of any authorized user with in the network trying to access the resource that is not dedicated for its use. Thus inside accidental malicious attacks could be controlled by efficient implementation of this feature.

8. *Detection*

This feature ensures control over unauthorized access into the security system, thus breaking/ hacking the system security.The implementation of this feature provides the functionality of detecting the attacks being launched from outside the network. As CAD security system provides protection to unauthorized access by the usage of access control lists so the monitoring module validate the access of the resources and data on network by detecting the combinations allowed by access control lists and in case an unknown combination is detected an alarm will be generated by the monitoring module for CAD security system.

9. *Automated Prevention*

This feature ensures the immediate control over response system once detection of a threat is announced. CAD security system provides automated prevention to unauthorized access by blocking the access of malicious user to the resources on the network. According to the analysis done on security breach incidents database maximum numbers of records are related to outside attacks. As breach methods like hacking are launched from outside the network and target data resources, so CAD security system provides encryption mechanisms to protect data such that an encrypted data will be not reveal the actual information. In addition to encryption techniques this system controls access by using access control lists that are controlled centrally and consist of access rights for every authorized user, so in case an authorize user tries to access a protected resource or data on the network that the user is not authorized to access, such an access will automatically be denied by the system.

10. Automated Monitoring

This feature ensures that the system is being monitored continuously without any delay, such that the immediate detection and prevention procedures are executed before a threat becomes a problem.CAD security system provides automated monitoring by implementation of a state based monitoring module that keeps record of active states of the users registered within the network.

11. Change Control Process

This feature ensures effective management of authorized changes when ever required in a system. Addition of this feature results in a configurable system where the roles of authorized users could be re-defined and the system can be easily updated for monitoring of new threats.

12. Data Backup and Event logging

According to the findings of analysis 95% of times data was not recovered after data breach incidents. CAD Security System supports data backup mechanism and event logging for purpose of data recovery in case of data loss, where system event logging would also help in tracking any loss of data in case a data breach attack is launched.

XV. Findings

(1) A total of 334 million people have been affected by the data breach incidents during years 2006 to 2009. There is an increase in number of people getting affected by data breach incidents recorded in year 2009, as the percentage ratio for people getting affected in 2009 is 41% highest in comparison to the records of years 2006 to 2008.

(2) In four different business domains mentioned in the records, Private businesses are the most affected by data breach incidents with a percentage ratio 52% as the highest percentage of incidents reported from this domain in comparison to other domains. According to statistics about 274 million out of the total number of people affected are related to Private business domain. It shows Private business domains are more vulnerable to data breach attacks as compared to other business domains.

(3) Data breach due to stolen laptops has the highest count of 441 out of 1663 records, showing that most of the times a major reason of data being compromised or leaked because of unprotected confidential data available on stolen laptops.

(4) The Breach type Hacking with a count of 299 is recorded as the second highest method of data breach attacks after stolen laptops, but according to the statistics observed after comparison of total affected versus breach types show that method of hacking has affected about 255 million people and is ranked as highest among other breach methods. This shows that a security system should have control over malicious and ambiguous events occurring within the network with features of active monitoring of resources under observation.

(5) Out of 1663 incidents reported 914 were related to Social Security Numbers, which is about 55% in total. According to statistics the attacks related with data type of Social Security Numbers have been ranked as high frequency reported incidents.

(6) In comparison of attacks on Business types and data types, it was observed that attacks on the private businesses targeted Social Security Numbers data and Financial data with frequencies of 345 and 295 (i.e. highest in comparison to frequencies of other data types) respectively.

(7) Although the highest number of attacks launched targeted Social Security Numbers data but in comparison of total people affected with Data types shows that the attack on financial data type has affected about 77% of 334 million people.

(8) Analysis of sources of attacks (i.e. Attack types) shows that out of 1663 incidents recorded 1113 are recorded as outside attacks. According to the comparison of total affected versus attack types outsider attacks have affected 90% of 334 million people resulting in a major threat of data security affecting a large amount of people.

(9) Out of 1663 incidents analyzed data lost in 95% of data breach incidents data was not recovered which shows that the features of data logging and feedback are missing from most security systems. This shows the requirement of efficient data recovery mechanisms to be supported by the security systems.

(10) Highest frequency outside attacks have affected all four types of business domains where a maximum of 473 incidents out of 1663 records used stolen laptops to compromise data and 281 incidents out of 1663 records outside attacks used method of hacking to launch data breach attack. This shows that a security system must support the features of threat detection, prevention and automated monitoring.

XVI. Conclusions

This paper aims to provide a detailed analysis of data breach incidents database and identify the real world problem in order to propose a solution that should be able to address the problems being currently faced by the industry and affecting millions of people. The identified core problems include threats to data confidentiality, unauthorized access of resources on the network, control over accidental exposure of data by authorized resource and lack of features for data recovery. In this paper a solution to these identified problems is proposed. The proposed Centralized Access Distribution security system is based on the role based access control model for basic data security features and in addition to the usage of this traditional model, nine components for achieving data security have been added to the CAD security system. Each component individually solves the problems identified after data breach records analysis and enhances the security on a CAD based system. The implementation of this CAD security system would help in protecting a networked business environment in terms of data confidentiality, data integrity, detection, prevention and monitoring of data breach attacks, easy configuration of system to protect against new threats and finally the event logging mechanism helpful in monitoring malicious activities going on the network.

XVII. REFERENCES

1) Xiaoni Liu Luyan Chen Cuiqin Duan, "Access control in Network Management System," Proceedings of 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS), 2009, Vol. 1, pp. 227 – 230.
2) Cheng Zang Zhongdong Huang Gang Chen Jinxiang Dong, "A Network Access Control Architecture Using State-Transfer-Based Dynamic Policy," In CSCWD '06: Proceedings of 10th International Conference on Computer Supported Cooperative Work in Design, 2006, pp. 1 – 5.
3) M.J.Moyer and M.Ahmad, "Generalized role-based access control," In ICDS '01: Proceedings of the 21st International Conference on Distributed Computing Systems, pp. 319-398.
4) Shih-Chien Chou, "Embedding role-based access control model in object oriented systems to protect privacy," The Journal of Systems and Software, Vol. 71, pp 143 – 161, 2002.
5) Bindiganavale, V. Jinsong Ouyang, "Role Based Access Control in Enterprise Application – Security Administration and User Management," Proceedings of IEEE International Conference on Information Reuse and Integration, 2006, pp. 111 – 116.
6) Wu Kehe Zhang Tong Li Wei Ma Gang, "Security Model Based on Network Business Security," In ICCTD'09: Proceedings of International Conference on Computer Technology and Development, 2009, Vol. 1, pp. 577 – 580.
7) Shucheng Yu Kui Ren Wenjing Lou, "Attribute-Based Content Distribution with Hidden Policy," In NPSec 2008: 4th Workshop on Secure Network Protocols, 2008, pp. 39 – 44.
8) G.A.S. Torrellas D.V. Cruz, "Security in PKI-based Networking Enviornment: AMulti-Agent Architecture for Distributed Security Management System & Control," In ICCC 2004: Proceedings of Second IEEE International Conference on Computational Cybernetics, 200, pp. 183 – 188.
9) Onno, S. Thomson R&D, Security Labs, Cesson-Sevigne, "A Federated Physical and Logical Access Control Enforcement Model," In ARES 08: Proceedings of Third IEEE International Conference on Availability, Reliability and Security, 2008, pp. 683 – 692.
10) Bao-Chyuan Guan Ping Wang Chen, S.-J. Chang, R.-I, "An Extended Object-Oriented Security Model For High Secure Office Environment," Proceedings of IEEE 37th Annual 2003 International Carnahan Conference on Security Technology, 2003, pp. 57 – 61.
11) Anwar, M.M. Zafar, M.F. Ahmed, Z, "A Proposed Preventive Information Security System," In ICEE '07: Proceedings of International Conference on Electrical Engineering, 2007, pp. 1 – 6.
12) Open Security Foundation DataLossDB website http://datalossdb.org/ breach
13) Wilde, E. Nabholz, N, "Access Control for Shared Resources," Proceedings of International Conference on Computational Intelligence for Modeling, Control and Automation, 2005, Vol. 1, pp. 256 – 250.
14) Gopinath, K, "Access Control in Communication Systems," Comsware 06: Proceddings of First International Conference on Communication System Software and Middleware, 2006, pp. 1 – 8.
15) Yue Zhang Joshi, J.B.D, "Temporal UAS: Supporting Efficient RBAC Authorization in Presence of the Temporal Role Hierarchy," EUC '08: Proceedings of IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, 2008, Vol. 2, pp. 264 – 271.
16) Harn, L. Lin, H.-Y, "Integration of user authentication and access control," IEE Proceedings of Computers and Digital Techniques, 2005, Vol. 139, Issue 2, pp. 139 – 143.
17) Yi Deng Jiacun Wang Tsai, J.J.P. Beznosov, K. Sch, "An Approach for Modeling and Analysis of Security System Architectures,"IEEE Transactions on Knowledge and Data Engineering, 2003, Vol. 15, Issue 5, pp. 1099 – 1119.
18) Yeu-Pong Lai, Po-Lun Hsia, "Using the vulnerability information of computer systems to improve the network security," Computer Communications, 2007, Vol. 30, Issue 9, pp. 2032 – 2047.
19) Denning, D.E, "An Intrusion-Detection Model," IEEE Transactions on Software Engineering, 2006, Vol. SE-13, Issue 2, pp. 222 – 232.
20) Felicia A. Durán, Stephen H. Conrad, Gregory N. Conrad, DavidP. Duggan, and E. Bruce Held, "Building a System for Insider Security," IEEE Security and Privacy, 2009, Vol. 7, Issue 6, pp. 30 – 39.

Network Design Problem Using Genetic Algorithm-An Empirical Study on Mutation Operator

GJCST Classification
C.2.1

Anand Kumar[1], Dr. N. N. Jani[2]

***Abstract*-This paper presents an influence of mutation operator in genetic algorithm for small to large network design problem. A network design problem for this paper falls under the network topology category which is a minimum spanning tree with various types of constraint which makes it NP-hard problem. Mutation operator plays an important role in genetic algorithm approach. Since many researchers have tried to solve this problem for small to mid size, we have explored the use of genetic algorithm with various mutation functions with modification but without changing the nature of genetic algorithm. Various mutation functions have been developed here as per the requirement of the problem and applied with the various size of network. In this paper we have tried to show that how mutation functions affects the performance of genetic algorithm and also shown that GA is an alternative solution for this NP-hard problem.**

Keywords-Genetic Algorithm, Network design, Mutation Operator Minimum spanning tree.

I. INTRODUCTION

In genetic algorithms of computing, mutation is a genetic operator used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next. It is analogous to biological mutation. The classic example of a mutation operator involves a probability that an arbitrary bit in a genetic sequence will be changed from its original state. A common method of implementing the mutation operator involves generating a random variable for each bit in a sequence. This random variable tells whether or not a particular bit will be modified. This mutation procedure, based on the biological point mutation, is called single point mutation. Other types are inversion and floating point mutation. When the gene encoding is restrictive as in permutation problems, mutations are swaps, inversions and scrambles.The purpose of mutation in GAs is preserving and introducing diversity. Mutation should allow the algorithm to avoid local minima by preventing the population of chromosomes from becoming too similar to each other, thus slowing or even stopping evolution. This reasoning also explains the fact that most GA systems avoid only taking the fitness of the population in generating the next but rather a random (or semi-random) selection with a weighting toward those that are fitter. There are many mutation schemes for genetic algorithms (Gas) each with different characteristics. Since the nature of genetic algorithm is very uncertain, various mutation operators can be used to derive optimal result. This paper presents the influence of various types of mutation operators with various size of network and it is the extension of the research work Network design problem [6][7]. This problem is one of the hardest problems in NP-hard category. There are no traditional methods available to solve this problem. A genetic algorithm approach to design the network is one of the ultimate solutions because traditional heuristics has the limited success. Researchers in operation research have examined this problem under the broad category of minimum cost flow problem‘ [1]. A simple GA approach is applied by many researchers [2],[3],[4] but in this paper we have shown the influence of mutation function in genetic algorithm. Genetic Algorithms are being used extensively in optimization problem as an alternative to traditional heuristics. It is an appealing idea that the natural concepts of evolution may be borrowed for use as a computational optimization technique, which is based on the principle Survival of the fittest” given by "Darvin". We have tried to show that the influence of mutation function and the little variation in genetic algorithm approach is very effective.

1. *Network Design*

In this paper network design is considered as network topology which is a spanning tree consists of various nodes considered as vertex. A tree is a connected graph containing no cycles. A tree of a general undirected graph G = (V,E) with a node (or vertex) set V and edge set E is a connected subgraph T = (V‘,E‘) containing no cycles with (n-1) edges where n is total no of node.In this study undirected networks are considered with the weight (distance) associated with each node. For a given connected, undirected graph G with n nodes, a minimum spanning tree T is a sub graph of a G that connects all of G‘s nodes and contains no cycles [5]. When every edge (i, j) is associated with a distance cij , a minimum spanning tree is a spanning tree of the smallest possible total edge cost

$$C = \Sigma\, c_{ij}$$

Where (i, j). $\in$ T

About[1]- Department of Master of Computer Applications AMC Engineering College, Bangalore kumaranandkumar@gmail.com
About[2]- Faculty of Computer Studies Kadi Sarva Vishwavidyalya, Gandhinagar drnnjanicsd@gmail.com

2. *Genetic Algorithm*

Genetic algorithms (GA) is a powerful, robust search and optimization tool, which work on the natural concept of evolution, based on natural genetics and natural selection..

3. *Work flow of GA*

1) Initialisation of parent population.
2) Evaluation
a) Self loop check
b) Isolated node or edge check
c) Cycle check
d) Store the best result
3) Selection of child population
4) Apply Crossover/ Recombination
5) Evaluation
6) Replace the result if it is better than previously stored.
7) Apply Mutation
8) Evaluation
9) Replace the result if it is better than previously stored.
10) Go to step 3 until termination criteria satisfies

II. Network Design Problem Presentation and its Solution Using Genetic Algorithm Approach

The Network design problem is considered as a unidirectional graph and represented with the help of adjacency matrix. Parent population in the form of chromosome is generated randomly according to the size of network. Number of gene in a chromosome is equal to number of node in a network. The total number of chromosome may vary and it is based on user input. Here a chromosome is generated for a 10 node network. The association between nodes is considered between positions to position.

node	1	2	3	4	5	6	7	8	9	10

chromosome	2	10	4	9	6	7	5	9	8	3

The logic behind association is that, the node [1] is connected with node 2; node [2] is connected with 10 and so on.

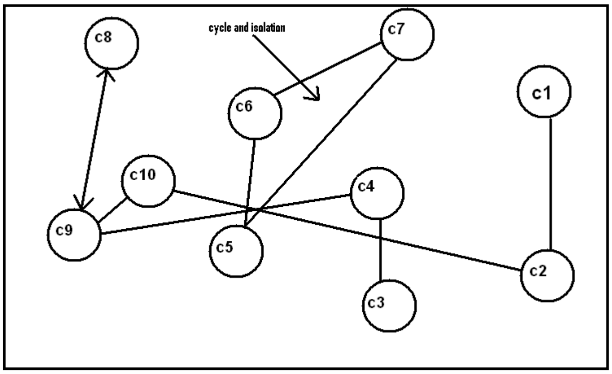


Figure 7

From fig-1 it is clear that this is not a spanning tree because of the isolated circle. Similarly with some other randomly generated chromosome, some other problems have been observed.

By observing these problems it has been concluded that there are three main reasons for illegal chromosome:

i. Self loop
ii. Cycle and
iii. Isolated node or edge.

1. Evaluation

By observing these problems, fitness functions have been developed [6] to evaluate these chromosomes. On the basis of these fitness functions, fitness points are given to the chromosomes and on the basis of these fitness points chromosomes are selected as a child population for next generation. Following fitness functions have been developed to evaluate chromosomes.

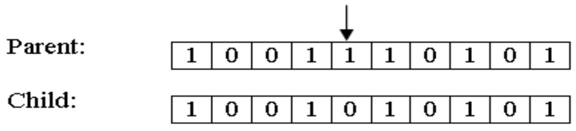
a) Self Loop
b) Isolated node or edge
c) Cycle

III. Mutation

This is the main part of this paper. Mutation is a background operator which produces spontaneous random changes in various chromosomes. A simple way to achieve mutation would be to alter one or more genes. In GA, mutation serves the crucial role of either (a) replacing the genes lost from the population during the selection process so that they can be tried in a new context or (b) providing the genes that were not present in the initial population. The mutation probability (denoted by Pm) is defined as the percentage of the total number of genes in the population. The mutation probability controls the probability with which new genes are introduced into the population for trial. If it is too low, many genes that would have been useful are never tried out, while if it is too high, there will be much random perturbation, the offspring will start losing their resemblance to the parents, and the algorithm [8]will lose the ability to learn from the history of the search. Up to now, several mutation operators have been proposed for real numbers encoding, which can roughly be put into four classes as crossover can be classified. Random mutation operators such as uniform mutation, boundary mutation, and plain mutation, belong to the conventional mutation operators, which simply replace a gene with a randomly selected real number with a specified range. Dynamic mutation (non uniform mutation) is designed for fine-tuning capabilities aimed at achieving high precision, which is classified as the arithmetical mutation operator. Directional mutation operator is a kind of direction-based mutation, which uses the gradient expansion of objective function. The direction can be given randomly as a free direction to avoid the chromosomes jamming into a corner. If the chromosome is near the boundary, the mutation direction given by some criteria might point toward the close boundary, and then jamming could occur. Several mutation operators for integer encoding have been proposed.

Inversion mutation selects two positions within a chromosome at random and then inverts the substring between these two positions.

- Insertion mutation selects a gene at random and inserts it in a random position.
- Displacement mutation selects a substring of genes at random and inserts it in a random position. Therefore, insertion can be viewed as a special case of displacement. Reciprocal exchange mutation selects two positions random and then swaps the genes on the positions.



Parent:	1	0	0	1	1	1	0	1	0	1
Child:	1	0	0	1	0	1	0	1	0	1

Figure 2

Following mutation operators have been developed and tested for different size of network.

Following data structure have been used for all these developed functions.

Chromosomes: it is a matrix of size N X N to store N randomly generated chromosomes. After the fitness function, fitness point for each chromosome is stored in its last column. Subscript starts from 1.

S (1): it holds the no of row of matrix chromosome

S(2): it holds the no of column of matrix chromosome

Fitness: it is an array which holds the fitness value for each chromosome.

1. *Mutation-I*

This mutation operator mutates only those chromosomes which does not have the maximum fitness. The logic applied behind this function is to simply find the chromosome and change its value with its position. If first chromosome is selected then its first place will be replaced by maximum number where maximum number is equal to number of node. Similarly if second unfit chromosome is selected then its second position will be replaced by maximum number-1 and so on.

```
Mutation1(chromosome)
Begin
Set k=1;
for i=1 to row do
if(chromosome(i, col) not equal to
maximum fitness)
      new_chromosome(i,i) = (col-i);
     end
end
for i=1 to row
    for j=1 to col-1 do
mutated_chromosome(i,j)=new_chromoso
me(i,j);
    end
end
End
```

2. *MutationII*

This mutation operator mutates only those chromosomes which does not have the maximum fitness value. Mutation is done to remove self loop. If the locus and allele both have the same vlaue, than this value is replaced by (position + 1). This function is also working as the repairing of chromosome.

```
mutationII(chromosome)
Begin
set k=1;
for i=1 to row do
     if(chromosome(i, col) not equal
to maximum fitness)
          for j=1 to col-1 do
              if(chromosome(i,j) ==
j)
                  if(j equal to col-
1)

new_chromosome(i,j) = j-1;
                  else

new_chromosome(i,j) = j+1;
                  end
               end
          end
     end
end
for i=1 to row do
    for j=1 to col-1
        mutated_chromosome(i,j) =
new_chromosome(i,j);
    end
```

```
end
End
```

3. *Random Mutation*

This mutation operator mutates only those chromosomes which does not have the maximum fitness value.Mutation is done by selecting a random position and replace its value with random number. It is considered that no self loop could form at the time of replacement.

```
Random_mutation(chromosome)
Begin
set k=1;
for i=1 to row do
     if(chromosome(i, col) not equal
to maximum fitness)
           posi = randomly generated
number within limit;
           val =  randomly generated
number within limit;
             if(posi equal to 0)
                 posi=1;
             end
             if(val equal to  0)
                 val=1;
             end
if((posi equal to val)AND (posi ==
col-1))
     chromosome(i,posi) = val-1;
             else
            chromosome(i,posi) =
val;
             end
        end
     end
for i=1 to row do
    for j=1 to col-1
        mutated_chromosome(i,j) =
new_chromosome(i,j);
    end
end
End
```

4. *Swap Mutation*

This mutation operator swaps two random position of each of the chromosomes .

If the randomly generated positions are 3 and 7.

			↓				↓			
Chromosom	5	1	4	9	8	2	1	3	10	1

After mutation-

Chromosom	5	1	1	9	8	2	4	3	10	1

```
Swap_mutation(new_chromosome)
Begin
Set k=1;
   for i=1 to row do

     p= randomly generated number
within the limit;
     q= randomly generated number
within the limit;
          temp = new_chromosome(i,p);
          new_chromosome(i,p) =
new_chromosome(i,q);
          new_chromosome(i,q) = temp;
   end
End
```

5. *Mutation Inversion*

This mutation operator inverts the genes between two random position for each of the chromosomes . For each chromosome there are different random position.
If the randomly generated positions are 2 and 8.

		↓						↓		
Chromosom	5	1	4	9	8	2	1	3	10	1

After mutation-

Chromosom	5	3	1	2	8	9	4	1	10	1

```
Mutation inversion(new_chromosome)
Begin
Set k=1;
   for i=1 to row do
     p= randomly generated number
within the limit;
     q= randomly generated number
within the limit;
     sort p,q
       for x = p to  q do
          temp = new_chromosome(i,x);
               new_chromosome(i,x) =
new_chromosome(i,q);
          new_chromosome(i,q) = temp;
          decrement q by - 1;
          if (x == q) || (x > q)
              break;
          end
         end
```

```
  end
End
```

6. *Mutation Insertion*

This mutation operator inserts one gene with another gene by displacing other genes. Two random positions are generated to denote two gene, then one random place gene is inserted with the another random place gene. Other inbetween genes are shifted. For each chromosome there are different random position.
If the randomly generated positions are 2 and 8.

↓ ↓

Chromosom	5	1	4	9	8	2	1	3	10	1

After mutation-

↓ →

Chromosom	5	1	3	4	9	8	2	1	10	1

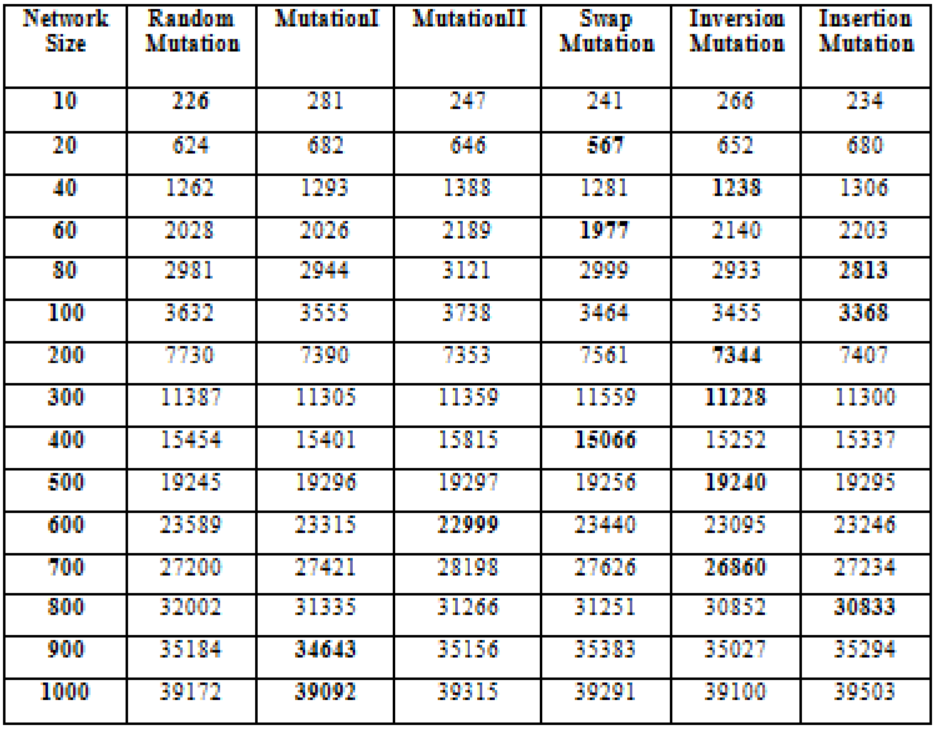
```
mutation_insertion(new_chromosome)
Begin
 Set k=1;
   for i=1 to row do
    p= randomly generated number
within the limit;
    q= randomly generated number
within the limit;
    sort p,q
       temp =  new_chromosome(i,
q);
        if(p not equal to  q)
            x = q-1;
  While (x greater than equal to
p+1)
                 new_chromosome(i,x+1) =
new_chromosome(i,x);
                  decrement x by -1;
          end
       new_chromosome(i,p+1) = temp;
        end
end
End
```

Crossover/Recombination
Chromosomes have been done on a single point.

IV. EXPERIMENTAL RESULT

Experiment based on Mutation Operator for small to large size network. The experiment is done in MATLAB R2008a version 7.6.0.324.
Following parameters have been considered:

Population size	: 100
No of Generations	: 100
Selection	: Roulette Wheel Selection
Crossover	: Uniform



Network Size	Random Mutation	MutationI	MutationII	Swap Mutation	Inversion Mutation	Insertion Mutation
10	226	281	247	241	266	234
20	624	682	646	**567**	652	680
40	1262	1293	1388	1281	**1238**	1306
60	2028	2026	2189	**1977**	2140	2203
80	2981	2944	3121	2999	2933	**2813**
100	3632	3555	3738	3464	3455	**3368**
200	7730	7390	7353	7561	**7344**	7407
300	11387	11305	11359	11559	**11228**	11300
400	15454	15401	15815	**15066**	15252	15337
500	19245	19296	19297	19256	**19240**	19295
600	23589	23315	**22999**	23440	23095	23246
700	27200	27421	28198	27626	**26860**	27234
800	32002	31335	31266	31251	30852	**30833**
900	35184	**34643**	35156	35383	35027	35294
1000	39172	**39092**	39315	39291	39100	39503

Table -1

Minimum Cost Of Network For Various Mutation Operators

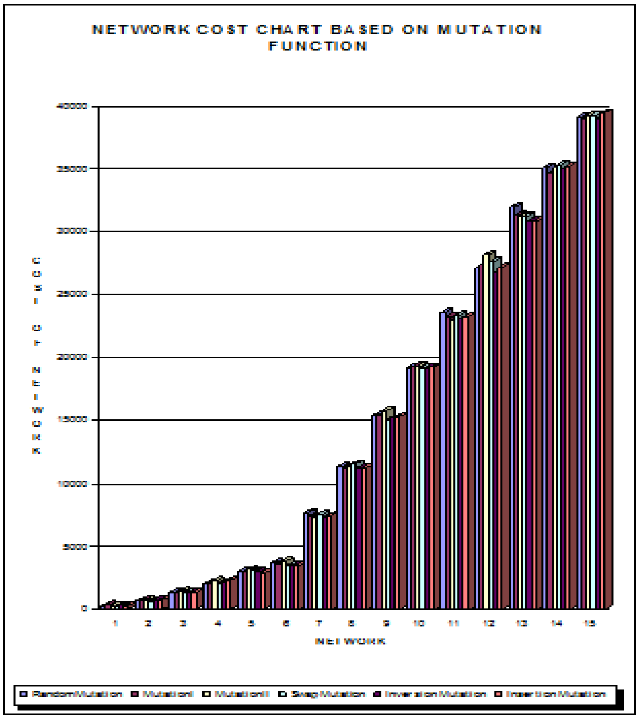


Figure 3

V. Conclusions

From the above experimental result of Table-1 and the chart shown in Figure-3, it is clear that, mutation operator is one of the important factor of genetic algorithm. From these six mutation function it is observed that insertion, inversion and swap mutation operator gives the better result.. This is the improved approach of evolutionary computing which gives the very positive result. We have described the importance of mutation operator. The effectiveness of the methodology however can be increased by applying the various genetic operators with variations of network size as the densely connected locations.

Acknowledgements

Our thanks to the reviewers for their cogent and insightful comments.

VI. References

1) Hamdy A. Taha. 2007. Operation Research An Introduction
2) S.K. Basu, 2005. Design Methods and Analysis of algorithms (PHI) ISBN : 81-203-2637-7
3) Melanie M. (1998). An Introduction to genetic Algorithm (PHI) ISBN 81-203-1385-5
4) Michael D. Vose. 1999. The simple genetic algorithm : (PHI) ISBN 61-203-2459-5
5) Narsingh Deo, 2000. Graph Theory with Applications to Engineering and Computer science: (PHI)
6) Anand Kumar, Dr. N.N.Jani, Proceeding of the International Conference on Mathematics and Computer Science ICMCS FEB 2010 Chennai.
7) Anand Kumar and Dr. N.N. Jani, Genetic Algorithm for Network Design Problem- An Empirical Study of Crossover operator with Generation and Population Variation" International Journal of Information Technology and Knowledge Management, ISSN: 0973-4414, Vol-III, Issue-I, June 2010
8) Donald Knuth. The Art of Computer Programming, Volume 1: Fundamental Algorithms, Third Edition. Addison-Wesley, 1997. ISBN 0-201-89683-4. Section 1.2.11: Asymptotic Representations, pp.107–123.

A Conceptual Study on Image Matching Techniques

GJCST Classification
F.2.2, I.4.0

Dr. Ekta Walia[1], Anu Suneja[2]

Abstract*-Matching Is The Technique To Find Existence Of A Pattern Within A Given Description. Image Matching Is An Important Application Required In The Field Of Image Processing. Images Are Represented As N-Dimensional Feature Vectors. Objects Of Same Class Possess Same Features And Those Objects Which Are From Different Class Possess Different Features. In The Image Matching Process, Features Are Used To Detect Whether Images Are Similar Or Not. Even We Can Find Whether Pattern Image Is A Subset Of Original Image Or Not. To Find Similarity Among Various Images Their Feature Vectors Are Matched. An Efficient Matching Technique Should Find Similarity Or Dissimilarity In Lesser Time Period. A Lot Of Matching Techniques Have Been Developed Till Today And Still Research For Developing An Optimized Matching Technique Is Going On. Most Commonly Used Matching Technique Is Nearest Neighborhood Technique. It Is An Important Technique Used In Applications Where Objects To Be Matched Are Represented As N-Dimensional Vectors. Other Matching Techniques Used Are Least Square Method, Coefficient Of Correlation Technique, Relational Graph Isomorphism Technique, Approximate Nearest Neighbor Technique And Matching Using Simulated Annealing Etc. All Of These Matching Techniques Have Their Own Advantages And Disadvantages. The Matching Technique Should Be Chosen Depending Upon The Application Area In Which It Is To Be Applied.

Keywords-Coefficient of Correlation, HSD, Image matching, K-NN, Nearest Neighborhood, Simulated Annealing, Sub Block Coding.

I. Introduction

Pattern matching is the technique to find existence of a pattern within an image. To localize a given pattern *'w'* in the image *'f'*, concept of mask is used. An image matrix of pattern *'w'* is the mask. This mask is placed over all possible pixel locations in the image *'f'* and contents of image mask and image *'f'* are compared. As a result of this comparison, a factor matching score is computed. If this matching score is greater than a predefined threshold value, the pattern *'w'* is said to be matched with some portion of image *'f'*. For the comparison of pattern *'w'* and image *'f'* various techniques have been proposed. the pattern *'w'* is said to be matched with some portion of image *'f'*. For the comparison of pattern *'w'* and image *'f'* various techniques have been proposed.

II. Matching Techniques

Image matching techniques are the techniques used to find existence of a pattern within a source image. Matching methods can be classified in two categories i.e. Area based matching techniques and feature based matching techniques. In Area based matching techniques, images are matched by numeric comparison of digital information in small sub arrays from each of the image. It includes methods such as Cross Correlation based matching technique, Least Square Region based technique and Simulated Annealing based matching techniques etc.In Feature based matching methods features of the image like edges, texture at different scales are extracted. Matching is performed with comparison based on characteristics of such extracted features. It includes methods such as Edge String based matching technique, Corner based matching technique and Texture region based matching technique etc [1].

1. *Coefficient of Correlation technique*

Yang,Y.,et al.[2] used least square method for matching process. In this technique, location (x_0,y_0)is pointed out find out in the image that minimizes the least square distance between original image *f(X,Y)* and pattern image *w(X,Y)*. The distance is calculated using equation

$$d^2(x,y)=\sum_{i=1}^{M}\sum_{j=1}^{N}\{f(X_i-x,Y_j-y)-w(X_i,Y_j)\}^2 \quad (1)$$

where *M X N* is the size of pattern *'w'*.
In cross correlation technique, derivative of least square is taken and pattern is matched by maximizing the second term of equation

$$d^2(x,y)=\sum_{i=1}^{M}\sum_{j=1}^{N}\{f(X_i-x,Y_j-y)-2f(X_i-x,Y_j-y)w(X_i,Y_j)-w(X_i,Y_j)\}^2 \quad (2)$$

But both Cross Correlation and Least Square methods get failed, if there is large variation in image intensity function. To remove this problem, *M.S. Sussman and G.A. Wright* proposed Correlation Coefficient technique for pattern matching. In Coefficient of Correlation technique, distance function is minimized in equation

About[1]- Professor and Head,Department of IT Maharishi Markandeshwar University Mullana, Haryana, India
About[2]- Lecturer, MMICT & BM Maharishi Markandeshwar University Mullana, Haryana, India E-Mail: anusuneja3@gmail.comContact: 094676-48895

$$d^2(x,y) =_j -y) - w^{'}(X_i,Y_j)\}^2 \qquad (3)$$

where $f^{'}(X,Y) = [f(X,Y) - \overline{f} / \sigma(f)]$

and $w^{'}(X,Y) = [h(X,Y) - \overline{h} - \sigma(h)]$

2. *Nearest neighborhood technique*

To implement Nearest Neighborhood technique, objects are first represented in the form of n-dimensional vectors. For such vectors, Euclidean distance is calculated to find similarity among various objects. Vectors having lesser distance have larger similarity. Euclidean distance in n-dimensional feature vector is distance between two vectors 'a' and 'b_i'.

Consider a = (a_1, a2,…..a_n) and b_i = (b_1,b_2,b_3…..b_n) then Euclidean distance is given as:

$$D_e(a,b_{ij}) = \sqrt{\sum_{j=0}^{n-1}(a_j - b_{ij})^2} \qquad (4)$$

Although Euclidean distance is commonly used measure of similarity, it is not the best method. More moderate approach that can be used to find similarity among vectors is to use the sum of absolute differences in feature vectors. It will save computational time and distance in such cases will be:

$$D(a,b_i) = \sum_{j=0}^{n-1}(a_i - b_{ij}) \qquad (5)$$

Euclidean distance method is translation, rotation and scaling invariant.Similar to Euclidean distance weighted and Mahalanobis distance methods are also used. In weighted Euclidean distance method weights w_j are assigned as weighted factor to show the importance of jth feature of vector [3]. Weighted Euclidean distance is defined as:

$$D(a_i,b_i) = \sqrt{\sum_{j=0}^{n-1} w_j(a_j - b_{ij})^2} \qquad (6)$$

In Mahalanobis distance method, statistical divergence properties of feature vector are used as weights. Distance in this method is given as:

$$D(a,b_i) = (a - b_i)^T \sum_i^{-1}(a - b_i) \qquad (7)$$

where summation over 'i' is variance-covariance matrix. Object having minimum Euclidean distance from 'a' is considered as nearest most similar object to 'a'.

3. *K-nearest neighbor technique*

The general version of nearest neighbor method is k-nearest neighbor method. In this method, nearest k neighbors are searched out rather than only one nearest neighbor.

A query in k-NN technique is defined as

Consider a vector 'a', and an integer 'k'. k-NN searches for k neighbors of vector 'a' according to distance between 'a' and kth neighbor. The result of k-NN query consists of k vectors such that:

$$||a\text{-}p|| \le ||a\text{-}q|| \qquad (8)$$

where $p \in R$

and $q \in DB\text{-}R$

R is result set of 'k' neighbors.

DB-R is the set of remaining points which are not among 'k' neighbors.

The drawback of k-NN method is its response time which is very large. The indexing matching methods have been proposed to overcome the problems of k-NN method.

To improve speed of K-NN image matching technique various approaches have been developed. For multidimensional feature space MIM technique has been proposed.

4. *MIM image matching method*

In MIM, feature space is partitioned into clusters and with the help of those partitions search is pruned. It has been observed that MIM works well for low dimensional feature space, even it works satisfactorily for high dimensions up to a threshold value[4][5][6][7][8].

To overcome dimensionality problem of MIM, a few other approaches have been developed.

a) The Dimensionality Reduction Technique
b) Approximate Nearest Neighbor Technique
c) Multiple Space Filling Curve Technique
d) Filter Based Technique

InDimensionality Reduction technique *G.Strang*[9] has suggested to condense most of the information into a few dimensions by applying SVD(singular value decomposition*)* technique. It will save time for indexing. According to *K.V.R. Kanth, D.Aggarwal and A. S*ingh although Dimensionality Reduction approach has solved dimensionality problem, but many other drawbacks have been observed. They are:-

a) Accuracy of query has been lost.
b) DR works well only if feature vectors are correlated.[9]

S.Arya[10] has discussed ANN technique to find 'k' approximate nearest neighbors in very short response time within an error bound 'e'.

Given a query vector 'q' and a distance error 'e' > 0 then, 'p' will be an ANN of 'q' such that for any other point 'p' in feature space

$$||\, q\text{-}p|| \le (1+e)||q\text{-}p'|| \qquad (9)$$

N.Megiddo and U. Shaft have discussed an approach in which n-dimensional space is reduced to 1-n space and gives linear ordering of all the points in the feature space. In multiple spaces filling curve approach n- dimensional space is arranged in 1- dimensional space according to mapping $R^d \rightarrow R^1$.

In this linear arrangement, nearer points on space-filling curve corresponds to nearer points in n-dimensional space. But the problem with this approach is that some nearest neighbors may be ignored in it [11] [12]. In filter based approach, *R.Weber,H.J. Schek and S.Blott* have reduced the range of vectors to be searched for pattern matching. In this technique only a few vectors are scanned during search of matching process. It returns exact K-NN of an object. Selected vectors with which object will be matched is extracted by filtering method.VA-file filtering approach has been discussed by R.Weber,H.J. *Schek and S.Blott* . VA-file divides feature space into 2^b rectangular cells. It allocate a unique b length bit string to each cell and approximate data points that fall in to cell by that bit string.K-NN queries are processed by scanning all approximations and by excluding majority of vectors from search based on these approximations. The problem with VA- file is that its performance converges to sequential scan and get worse as number of bits used for approximation get increased. To remove this problem *Guang-Ho, Xiaoming Zhu*[13] has proposed an *efficient indexing method for NN searches in high-Dimensional image database.* In LPC-file some additional information is stored which is independent of dimensionality.

5. *LPC-File indexing method*

An indexing matching method has been proposed by *Guang-Ho, Xiaoming Zhu*[13] to reduce the response time of k-NN technique. Indexing method proposed by them was named as LPC-file. LPC is for local polar coordinate file. LPC-file improves search speed even in high dimensional feature space [13]. LPC-file is a filter based approach for image matching. LPC-file approach is similar to VA-file, but it adds polar coordinates information of vector to the approximation. It is sufficient to use 3 bytes for polar coordinates, 2 bytes for radius and 1 byte for angle.Unlike MIM, where cells are organized in hierarchical manner, in LPC vector space is partitioned into rectangular cells and then these cells are approximated by their polar coordinates. In LPC-file for each vector 'p_i', where $i \in \{1, 2, 3------N\}$, an approximation *'a_i'* is found out. In next step vector *'P'* is represented using polar coordinates *$(r, \theta\ )$* in the cell in which *'P'* lies.

Thus *'P'* is represented as triplet *$a = (c, r, \theta\ )$*

Where c is approximation cell, *'r'* is radius and 'θ' is angle of *'P'*.

Complete LPC-file is an array of approximations of all the vectors. To find out K-NN only filtered vectors are stored in LPC- file are scanned.

On the basis of approximations of vector, bound on the distance between query point and vector is derived to restrict the search space between k-NN searches.

$$d_{min} = \ |\,p|^2 + |q|^2 - 2|p|\,|q|\cos|\,\theta_1 - \theta_2| \qquad (10)$$

and

$$d_{max} == \ |\,p|^2 + |q|^2 - 2|p|\,|q|\cos(\theta_1 - \theta_2) \qquad (11)$$

In filtering process, vectors are collected to form candidate set.For this collection, each vector's d_{min} and d_{max} is computed. If a vector is found where d_{min} exceeds the distance k-$NN^{distance}$(q) of kth NN encountered so far, then corresponding vector can be eliminated since k better candidates have already been found.

Consider a 5-dimensional vector space V= {orientation, x, y, scale, intensity}. In 5 dimensional vector 3 bits will be used for assigning bit string to each dimension.

According to LPC, we also store *'r'* and 'θ' . On the basis of value of *'r'* and 'θ' vectors are filtered.

6. *Image Matching By Simulated Annealing*

A number of matching techniques have been developed but the problem with such matching techniques is that they have very high response time for matching process. An optimized image matching technique has been developed by *Laurent Herault, Radu Horaud*[14]. This technique was based on simulated annealing, where firstly image is represented in the form of relational graph. Then a cost function is derived for the graph of the image. This cost function is optimized with the method of simulated annealing.To use simulated annealing method for image matching, description of image is represented in the form of relational graph. In this graph nodes represent features and arcs represent relation among these features. This relational graph is casted into optimization problem and such problem is solved using simulated annealing technique.For simulated annealing cost function is represented as quadratic function. It will help to calculate energy variation in annealing process. In physical annealing process, in order to reach at a low energy state, metal is heated up to high temperature and then is cooled down slowly.To apply simulated annealing for image graph, states, state transition, random generation of state transition and change in energy associated with state transition is explicitly defined [14]. Let *'a'* and *'b'* be the two graphs where *'b'* should be isomorphic to *'a'*. To optimize the matching process, isomorphism among *'a'* and *'b'* must minimize the equation

$$E = \sum_{s=1}^{S} \lambda_s E^s \qquad (12)$$

where *'E'* is cost function and *'S'* is the number of possible relationships in graph *'a'* and *'b'*and

$$E^s = \sum_{k=1}^{N} \sum_{l=1}^{N} (1 - 2a_{kl}).b_{\pi(k)\pi(l)} \qquad (13)$$

Where N= number of nodes in graph,

λ_s = weight assigned to each of relationship

Π =one-one correspondence between vertex of 'a' and vertex of 'b' which minimize the distance between two graphs.

7. *HSD (Histogram Based Similar Distance) based Matching Technique*

Boaming Shah, Fengying cui[15] presented HSD (Histogram Based Similar Distance) technique combined with ARPIH(Angular Radial Partitioning Intensity Histogram) matching technique to find number of matching points between source and target image. HSD provides high performance for geometric attacks like rotation and shearing. It gives better performance even in case of illumination change.

Using ARPIH technique, a strength histogram is constructed and considered as an image. In ARPIH descriptor, image is partitioned into 18 sub regions according to angle θ which are (π/3, 2π/3, π, 4π/3, 5π/3, 2π) and the ratio of radius 'r'.

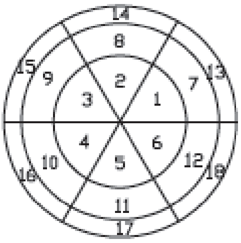


Fig 1: ARPIH image subregions

In ARPIH descriptor a two-dimension histogram is constructed which represents the pixel grayscale distribution in the image region and the geometry relationship between the sub regions. The x-axis of histogram is the serial number of sub region, and y axis is grayscale (0-255) which is evenly divided into 18 gray ranges. Then the pixels in every sub region are distributed into every gray range by its own grayscale.

HSD is based on MAD(Mean Absolute Difference Algorithm) and MLD technique. In these techniques rather than calculating all point's distance from one aggregate to another, distance between two corresponding points is taken as main similarity measure. Therefore, the similarity between every pair of corresponding points is calculated and then similarity is accumulated according to minimum difference to get distance between two images.

To explain it consider the template image as S(m, n), its size is M×N, the target image as I(u, v), its size is U×V. The position of template image in the target image is (i, j) , suppose S'(m , n)=I(i+m , j+n), d(i , j) denotes the distance function between the same size image windows, (i*, j*) denotes the optimal matching position, 'P' is the matching range, 'P' is defined as follows:

$$P=\{(i,j),0\le i\le\le U-M, 0\le j\le V-N\} \qquad (14)$$

The distance measurement function based on traditional mean absolute difference algorithm (MAD) is defined as follows:

$$d(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}R_{MAD}(S(m,n),S^{'}(m,n)) \qquad (15)$$

where

$$R_{MAD}(S(m,n),S^{'}(m,n)) =| S(m,n)-S^{'}(m,n)) \qquad (16)$$

The optimal matching position is

$$d(i^{*},j^{*})=\min\{d(i,j)\,|\,(i,j)\in P\} \qquad (17)$$

MLD is defined as follows [16]:

$$d(i,j)=\sum_{m=1}^{M}\sum_{n=1}^{N}R_{MAD}(S(m,n),S^{'}(m,n)) \qquad (18)$$

where

$$R_{MAD}(S(m,n),S^{'}(m,n))=\left\{\begin{array}{l}1,|\,S(m,n)-S^{'}(m,n)\,|\le T\\ 0, else\end{array}\right\} \qquad (19)$$

where d(i ,j)means similarity, so the optimal matching position is

$$d(i^{*},j^{*})=\max\{d(i,j)\,|\,(i,j)\in P\} \qquad (20)$$

The difference between the two algorithms is that the former computes the sum of the entire pixel's grayscale absolute difference, while latter one only computes the number of the similar points. To perform matching only the similar number of points are considered between the template image and target image, to measure the similarity degree, and at the same time it discards those points that have more differences with the template.In HSD matching technique, histograms for both images template and target are drawn and two images are considered similar if they satisfy following conditions:

$$D_{HSD}\ge T_{1} \qquad (21)$$

$$D_{HSD}=\sum_{m=1}^{M}\sum_{n=1}^{N}R_{HSD}(H(m,n),H^{'}(m,n)) \qquad (22)$$

$$R_{HSD}(H(m,n),H^{'}(m,n))=\left\{\begin{array}{l}1,|\,H(m,n)-H^{'}(m,n))\,|\le T_{2}\\ 0, else\end{array}\right\} \qquad (22)$$

where, T1 and T2 are the threshold values which are predefined.

Steps followed during HSD are as:

1) Find ARPIH of the template image.
2) Select the sub-region from the top left corner of the target image in the same size with the template image, and find its ARPIH.
3) Match the two histogram according to HSD technique
4) Glide the template image on the target one, and search the sub-region with the same size as template image, then get its ARPIH.

5) Repeat above until to finish a whole scan for target image, the matching position is the area which has the maximal DHSD value.

8. *Sub Block coding based matching technique*

A template based matching method is discussed by *Yuping FENG, Shi LI, and Ming DI* [17]. This method combines local gray value encoding matching technique and phase correlation technique. Here matching is divided into two parts: rough matching and fine matching. In Rough matching image is divided into certain blocks called R-blocks, and sum of gray value of each R-block pixel is calculated. After that R-blocks are encoded according to gray value distribution of R-block with neighboring R-blocks and matching is performed between template and each search sub image. The detailed description of this method is as:

A. *Rough matching*

An image of size N X N is divided into some k × k size non overlapping blocks called R-block. Each R-block has eight neighborhoods and four D neighborhoods, and they have the following relations:

$$D1 = R1 \cup R2 \cup R4 \cup R5$$

$$D2 = R2 \cup R3 \cup R5 \cup R6$$

$$D3 = R4 \cup R5 \cup R7 \cup R8$$

$$D4 = R5 \cup R6 \cup R8 \cup R9$$

D Neighborhoods are sorted on the basis of sum of their pixel's gray value. There are 24 (4!) kinds of possibility for sorting the gray value sum of each R block pixel in every D-neighborhood. Every possible sorting result can be represented by five bits binary code, that is P (Dj) belongs to {00000, 00001… 10111}. Each R-block has four D-neighborhood, each D-neighborhood has one five bits binary code. The Ri block coding is to connect the adjacent D-neighborhood code, and obtain twenty bits binary code, which is:

F(Ri)=P(D1)P(D2)P(D3)P(D4) (24)

where F (Ri) is called R-block's code. Coding features of an image are made up of all R-block's codes. Through encoding, the content of the image is present as R-block's codes which indicate the different spatial gray value distribution of the image. Similarity among images is found as: more same feature codes images have, more similar they will be. It reduces the complexity and computation of matching. The more sub-blocks are divided, the better the content of image is described deviously, but it increases the encoding and matching computation. To improve searching speed scanning point by point is replaced with scanning of pixels at certain steps. This search strategy greatly saved calculation time.

The R-block's code set is expressed with (N/ k) orders square matrix A□ T and A□ S i ,j□ that is called characteristic coding matrix, 'k' is the size of R-block and N×N is the size of 'T'. Matching takes the number of the same feature code as similarity measure. The more same feature codes they have, the more similar regions they have. In Rough matching process the template and search image are divided into some non overlapping R-blocks, then characteristic coding matrix of the template and all R-block's code of the search image is calculated; after that search image is scanned by step, the coding matrix of template and each search sub-image is compared with to get the number of the same element recorded, say 'w'. Last, the location of the largest 'w', 0(io, jo)is the final coarse matching result. During the process of scanning, if the value of 'w' is greater than a certain threshold, the matching is interrupted.

B. *Fine Matching Using Phase Correlation*

In rough matching, the scanning process is performed by a certain step, where the template and search sub-image may be not completely overlapped. Therefore, rough matching result may not represent the correct matching. Due to this, precision matching amendment using phase correlation is adopted after rough matching.In Fine matching temporary matrix from image is taken on the basis of result of rough matching. After that phase correlation between temporary matrix and template is calculated. Based on cutting position and the matching result translation factor is obtained for matched template. On the basis of this phase correlation factor final matching is performed. Phase correlation method [18] based on Fourier transform is used for estimating the translation by phase relationship; it is not impacted by the different image content. The linear transformation of pixel grey value and image noise mainly effect amplitude in frequency domain but not its phase. Phase correlation has higher matching accuracy because of sharp correlation peak, and also has the stability of small-angle rotation. Assuming that the following translation relations between 'g1' and 'g2' images are given:

$$g_1(x, y) = g_2(x - x_o, y - y_o) \qquad (25)$$

The Fourier transform for above equation is as:

$$G_1(u, v) = G_2(u, v)e^{-j2\pi(ux_o + vy_o)} \qquad (26)$$

The cross-power spectrum for g1 and g2 is as:

$$\frac{G_1(u, v)G_2^*(u, v)}{|G_1(u, v)G_2^*(u, v)|} = e^{-j2\pi(ux_o + vy_o)} \qquad (27)$$

The phase correlation function is that:

$$corr(x, y) = F^{-1}(e^{-j2\pi(ux_o + vy_o)}) = \delta(x - x_o, y - y_o) \qquad (28)$$

where $\delta(x - x_o, y - y_o)$ is a pulse peak function. The biggest peak position is the translation(x0, y0). The temporary matrix whose size is the same as template size is cut from search image according to rough matching point (i0, j0). The phase correlation translation (x0, y0) between temporary matrix and T is used to modify the rough match, as following:

$$(x, y) = (i_o + x_o, j_o + y_o) \qquad (29)$$

Here (x, y) are the finally matched coordinates.

III. Conclusion

In Image Processing applications matching is very important phase. For the application having vectors of low or medium dimensions, MIM, R* tree and SR tree etc. are perfectly affordable. As dimensions increases filtered based approaches should be used to shorten the response time of matching process. LPC- file and VA- file are filtered based matching methods. LPC-file outperforms VA-File and K-NN matching methods. Its response time and disk space consumption is very less as compared to other matching techniques. It works well for both random and skewed distributed vector space. To optimize matching process of images simulated annealing technique can also be preferred. A combination of local gray value encoding matching and phase correlation matching technique gives two times better performance than existing sub block coding based matching techniques. HSD is technique used for matching images affected by geometric attacks like rotation etc.

VII. References

1) Balleti, F. Guerra "Image matching for historical maps comparison," e-perimetron, Vol. -4 , No. 3, 2009, pp 180-186.
2) Yang, Y., et al., MRM, 957, 1996.
3) B.Chanda, D.Dutta Majumder, "Digital Image Processing and Analysis", PHI, pp 335-357
4) N. Beckmann, H.P. Kriegel, R. Schneider and B. Seeger, "The R* tree: An efficient and robust access method for points an rectangles", in Proc. ACM SIGMOD Int. Conf. Management of data , 1990,pp. 322-331.
5) S. Berhtold, D. A. Keim and H. P. Kriegel , " The X-tree: an index structure for high dimensional data" in proc. 22nd Int. Confe, very large databases, 1996, pp 28-39.
6) H. Cha and C. W. Chung, "A new indexing scheme for content based image retrieval", Multimedia Tools Applications, Vol. 6, no. 3, May 1998,pp 263-288.
7) T.M. Cover and P. E. Hart, "nearest neighbor pattern classification", IEEE Transactions Information Theory, Vol. IT-13, 1997,pp 21-27.
   A. Henrich, "The LSD-tree: an access structure for feature vectors", in proceeding 14th International conference data engineering, 1998, pp 362-369.
8) Strang , Linear Algebra and its applications , 2nd edition, New York: Academic, 1980.
9) S. Arya, D. M. Mount, N.S. Netanyahu, R. Silverman, And A.Y.Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," J. ACM, vol. 45, no. 6, 1998,pp 891-923.
10) N.Megiddo and U. Shaft, :Efficient nearest neighbor indexing based on collection of space filling curves," IBM Almaden research center, San Jose, CA, Tech., 1997.
11) Shepherd, X, Zhu and N. Meggiddo,' A fast indexing method for multidimensional nearest neighbor search," in proceeding IS&T/SPIE Conf. storage and Retrieval for image and video databases, 1999,pp 350-355.
12) Guang – Ho Cha, Xiaoming Zhu, Dragutin Petovic, Fellow, IEEE and Chin-Wan Chung ,"An Efficient Indexing method for nearest neighbor Searches in high dimensional Image databases", IEEE Transactions on Multimedia, Vol 4, No. 1 , March 2002,pp-76-87.
13) Laurent Herault, Radu Horaud, "Symbolic Image Matching by Simulated Annealing", pp-31-324.
14) Baoming Shan, Fengying Cui," Image Matching Based on Local Invariant Feature
15) And Histogram-based Similar Distance ", 2009 First International Workshop on Education Technology and Computer Science © 2009 IEEE
16) Yuntao Liao, Xianyi Ren, Guilin Zhang, etc, A New Image Matching Method Based on a Novel Pixel Distance Measurement [J], Infrared and Laser Engineering, 2001, 30(6), 418–421.
17) Yuping FENG, Shi LI," An Image Matching Algorithm Based on Sub-block Coding ", 2009 Second International Workshop on Computer Science and Engineering © 2009 IEEE.
18) H.WANG, L.S. WANG, J. G. GAO, and S. ZHANG, "Fast stitching algorithm of sequence images for auto rack girder detection", Optics and Precision Engineering , vol 16, no 3, March 2008,pp 531-536.

A Contribution to the Development of Counter Measures for Dead Dropping in Cyber Warfare

GJCST Classification
D.4.6, K.6.5

Nayani Sateesh

***Abstract* - In recent years, cyber warfare has become the major threat to world safety. Terrorists are using the internet as their weapon to do suicide attacks, hijackings, bomb blasting, attacking the networks etc to create the grate damage. In order to defend against future attacking, it is important to understand how they are making use of the internet services and hence to create the counter measures. In this paper, we are considering an internet service called email - which is used for dead dropping to do cyber warfare, and developing a framework to counter it.**

Keywords-cyberwarfare, dead drop, Autherisation, Authentication.

I. Introduction

Internet has become critically important to the financial viability of the national and global economy. Internet made this world as a global village and kept everything on the web in a single click distance. It makes the people to get connected regardless of their geographical areas and distances. internet provide various services to the users which includes e-mail, FTP, Telnet, Archie , GOPHER , finger, Usenet and Mailing list , WWW etc. These services facilitate the user to get connected to the remote users, resources and can share the knowledge. Apart from these, we are also witnessed the cyber terrorism by the terrorists who used internet as their weapon. One of such service used by them in the recent attacks is emails using the dead drop method. Dead drop is the method in which the users will share the username and password for the email account in which they will store and share the messages in the email drafts.

II. Problem under study

Permissions over the resources in the networks or internet are granted to the users with the help of authorization and authentication mechanisms.Authorization is the function of specifying access rights to resources, which are related to information security , network security etc. It involves defining the access policies Authentication is the process by which we can verify that someone is who they claim they are. This usually involves a username and a password, but can include any other method of demonstrating identity, such as a smart card, retina scan, voice recognition, or fingerprints. Once the username and password are shared, then any person can access the resources on the network. This pitfall has given strength to the attackers to share their credentials to access the resources over the networks who are geographically dispersed with the same credentials. This advantage has given a shape to dead dropping Dead drop is the method in which the users will share the username and password for the email account in which they will store the messages in the email drafts. Those who are shared the username and password can only login and read them, in which there is no scope to identify where the mail is generated or received as the contents are stored in drafts only. The identity of the sender or receiver is only known once the mail is sent over the internet. We can find the identity using the email headers using various email header analyzer tools. In this paper we are not focusing on these tools but can have a look at way we can analyze the headers to get the identity of sender and receiver.In the following section we will discuss on the framework that we are proposing and way it could work. This framework can be easily implemented from intranet to the internet level.

III. Related work

Whenever a cyber crime is done over the internet through emails, the identity of the accused is getting find out using the email headers only if and only if the mails are transmitted over the internet only.

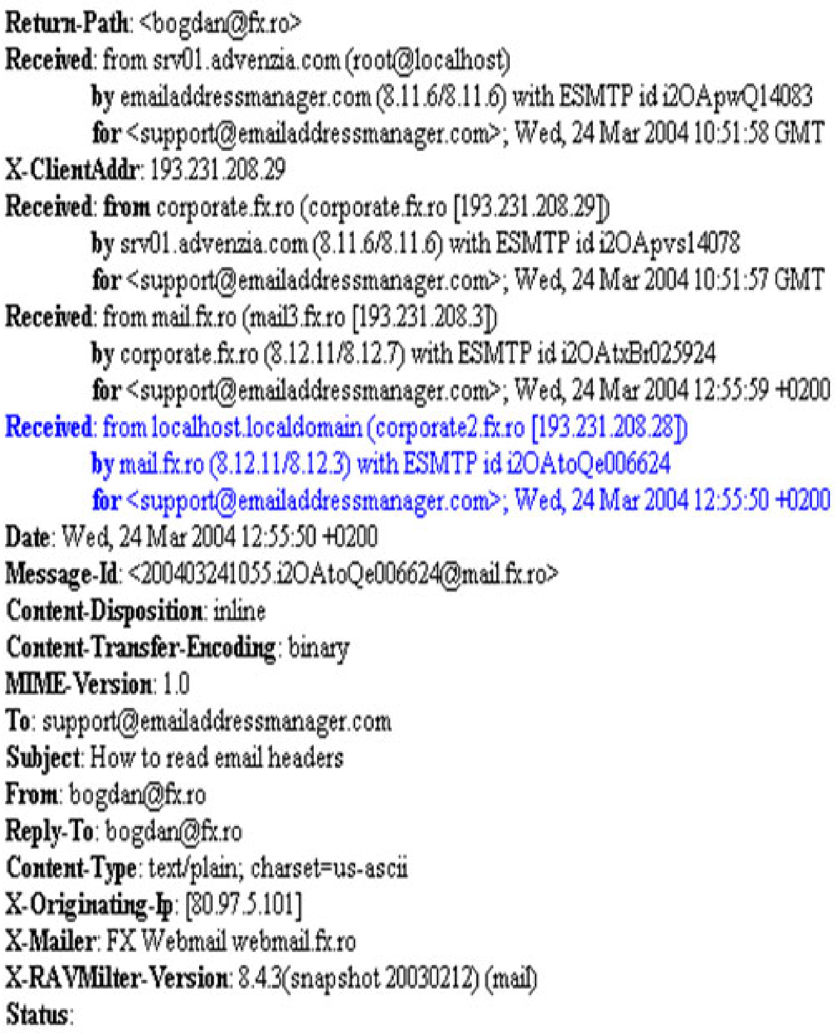
Email Header

The email header is the information that travels with every email, containing details about the sender, route and receiver. It includes the details such as who sent the email, when the email was sent, from where it was sent and how did it arrived and who is the receiver and when it was received.

Email headers interpretation

Let‘s take an example (we will ignore the header tags that do not give precise information about the sender). The following email was received by support@emailaddressmanager.com and we want to see who the sender is. Here is the email header of the message:

About- Hyderabad, Andhra Pradesh, India – 500081nayanisateesh@gmail.com



```
Return-Path: <bogdan@fx.ro>
Received: from srv01.advenzia.com (root@localhost)
        by emailaddressmanager.com (8.11.6/8.11.6) with ESMTP id i2OApwQ14083
        for <support@emailaddressmanager.com>; Wed, 24 Mar 2004 10:51:58 GMT
X-ClientAddr: 193.231.208.29
Received: from corporate.fx.ro (corporate.fx.ro [193.231.208.29])
        by srv01.advenzia.com (8.11.6/8.11.6) with ESMTP id i2OApvs14078
        for <support@emailaddressmanager.com>; Wed, 24 Mar 2004 10:51:57 GMT
Received: from mail.fx.ro (mail3.fx.ro [193.231.208.3])
        by corporate.fx.ro (8.12.11/8.12.7) with ESMTP id i2OAtxBr025924
        for <support@emailaddressmanager.com>; Wed, 24 Mar 2004 12:55:59 +0200
Received: from localhost.localdomain (corporate2.fx.ro [193.231.208.28])
        by mail.fx.ro (8.12.11/8.12.3) with ESMTP id i2OAtoQe006624
        for <support@emailaddressmanager.com>; Wed, 24 Mar 2004 12:55:50 +0200
Date: Wed, 24 Mar 2004 12:55:50 +0200
Message-Id: <200403241055.i2OAtoQe006624@mail.fx.ro>
Content-Disposition: inline
Content-Transfer-Encoding: binary
MIME-Version: 1.0
To: support@emailaddressmanager.com
Subject: How to read email headers
From: bogdan@fx.ro
Reply-To: bogdan@fx.ro
Content-Type: text/plain; charset=us-ascii
X-Originating-Ip: [80.97.5.101]
X-Mailer: FX Webmail webmail.fx.ro
X-RAVMilter-Version: 8.4.3(snapshot 20030212) (mail)
Status:
```

There are three paragraphs starting with the Received tag: each of them was added to the email header by email servers, as the email travelled from the sender to the receiver. Since our goal is to see who sent it, we only care about the last one (the blue lines). By reading the Receiving From tag, we can notice that the email was sent via corporate2.fx.ro, which is the ISP domain of the sender, using the IP 193.231.208.28. The email was sent using SMTP ("with ESMTP id") from the mail server called mail.fx.ro. Looking further into the message, you will see the tag called X-Originating-IP: this tag normally gives the real IP address of the sender. The X-Mailer tag says what email client was used to send the email (on our case, the email was sent using FX Webmail).

IV. Proposed Framework

We are proposing a framework to counter the dead drop using the MVC Architecture. The following figure shows the typical function of the each module.

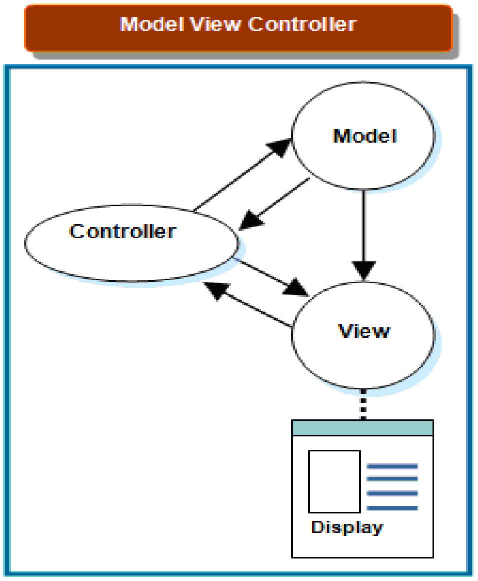


Fig: 4.1 – MVC Architecture

In MVC, View module deals with the user interfaces through which the user will get interact with the applications. Controller deals with the events and the invoking of the appropriate applications to provide the services. Model component deals with the underlying database. The controller module will be working an interface between the View and Model components in the MVC architecture In our proposed work, the user will be given an interface to login to the email service. The user will enter his user name and password. Once the user is logged in, we will take the IP address of the user through Server side Includes (SSI). Here is a typical syntax in javascript to get the IP address of the host.

```
<script language ="javascript">
var ip = '<!--#echo var="REMOTE_ADDR"-->';
</script>
```

Whenever an entry is made in the drafts or drafts were read, an event will be fired and the controller will send the mail content to the content analyzer where the mail contents will be analyzed. The content analyzer should be sophisticated with the text / data mining techniques such as classification, clustering, Bayesian classification etc.so that it will analyze the mail text or the attachment contents about the significance of threaten. The analyzer can be built-up with the capability to analyze the words, phrases, content type. If the content is significantly related to the cyber attack, then the controller will invoke an application saying that the draft is something related to malicious or threat and send to an appropriate authority as message or mail along with the IP address where the drafts entry is made or the draft is read.

V. Limitations

Dial-up users may have a different IP address each time they connect, and many other users may be behind proxies so that hundreds of machines will all report the same IP address

VI. Conclusion

Counter Measures for the Dead Dropping is still under research. The efficiency of this framework should be analyzed at the intranet lever before enhancing it to the internet mail applications.

VII. Future Work

We can also implement the biometric approach such as identifying the people by their typing patterns. Need to integrate the well-versed content analyzer algorithms for classifying the content correctly in email drafts.

VIII. Acknowledgments

I would like to thank the anonymous reviewers for their insightful feedback, which greatly improved the quality of this paper. I also thank one and all who helped me in making this paper published

IX. References

1) S. Sellke, N. Shroff, and S. Bagchi, Modeling and Automated Containment of Worms", IEEE Transactions on Dependable and Secure Computing, PP. 71-86, Vol. 5, No. 2, April-June 2008
2) ―How to read email headers" http://www.gradwell.com/support/howto/article/404
3) Behrouz A. Forouzan, Catherine Ann Coombs, Sophia Chung Fegan Data Communications and Networking
4) HTML Black Book - Steven Holzner
5) A Whitepaper on Effective Content Analysis for Email Inspection and Control", By Nemx Software Corporation, Canada.http://www.nemx.com/documents/Content%20Analysis%20Whitepaper.pdf

Bio Inspired Cross Layer Aware Network protocol for Cognitive Radio Networks

Vibhakar Pathak[1], Dr. Krishna Chandra Roy[2], Santosh Kumar Singh[3]

GJCST Classification C.2.1

***Abstract*-The reconfigurability and flexibility of cognitive radio heralds an opportunity for investigators and researcher community to reexamine how network layers protocols enhance quality of services(QoS) by interacting with lower layers of network services. Paper investigates enhancements of cognitive radio based computer networks. The enhancements are in form of better and agile network layer protocols ,which reconfigure itself as per need and change in physical network structure. Top operation like addressing, framing and error control are not modified . Super frame structure , flow control are reconfigured as per physical and data link parameter. Appropriate techniques are employed for better QoS for broad range of services over CRCN. The proposed Bio inspired Cross Layer Aware Network(BCLAN) Protocol is presented . The beauty of the protocol is that it uses ANT Colonization optimization which thought to be one of the best algorithm design strategy The protocol (BCLAN) is also tested for various services like VOIP,IM and FTP. The QoS found to be better than existing proactive and reactive routing protocols. The simulation was done on OMNET++ discrete simulation environment.**

Keywords-Cognitive Radio, Wireless system, Bio Inspired, Cross-layer aware, Network layer, QoS.

I. Introduction

Recent development in silicon technology leads to development of smart reprogrammable circuits Using which a new class of intelligent or “COGNITIVE” radios can be develop based on Software Defined Radio(SDR).Such radio based system would be capable of dynamic physical adaptation. In recent past development of cognitive radio hardware and software, especially at the physical layer has received considerable attention. The question how one can transform a set of cognitive radio into a cognitive network is less considered by research community. Cognitive radio or agile radio is a technology to choose a wide variety of radio parameters and protocol standard in adaptive manner on observed radio link and network conditions. The reconfigurability and flexibility of cognitive radio heralds an opportunity for investigators and researcher community to reexamine how network layers protocols enhance quality of services(QoS) by interacting with lower layers of network services. Present wireless protocols define reliable service parameter within layers of network protocols, which lacks optimization. The cognitive radio(CR) enables spectrum aware communication system by which any protocol can reconfigure itself for best performance . Such type of system is capable of handling cross layer parameter change and advice the network to change the system with view of minimum power consumption, lowering back off and reducing rate of drop packets, and hence updating utilization of network resource. In generalized case the cognitive radio is capable of adapting modulation of wave form, OSA (opportunistically spectrum access) ,MAC protocols , network protocols. The cognitive radio can make runtime change to protocols to avoid collisions by transmitting packets with minimum power utilized for hop to hop transfer. There have been many research work addressing physical layer agility of cognitive radio system based on OSA[2][3]. Our goal is to investigate Bio Inspired cross layer aware protocol design for network layer (BCLAN). The paper also tries to amalgamate bio inspired computing for higher order optimization. The paper proposes Bio Inspired Cross Layer Aware Network (BCLAN) layer protocol for Cognitive Radio Computer Network (CRCN). The protocol is based on ANT colonization and their hunt for food .The route towards food and there optimization by worker ant is utilized here. Extensive OMNET++ simulation shows that QoS is significantly enhanced by using BCLAN in combination with KR-MAC in fixed packet size .For broader range of network services a combination of DNA sequence alignment based spectrum sensing and CLA-AMAC are used .The result of above protocols are encouraging one. The rest of paper is organized as follows. Back ground and other related works is presented in section 2, section 3 presents the proposed BCLAN protocol .Simulation results and analysis are presented in section 4 , conclusion and future directions are presented in section 5.

II. Back ground and Related works

S. Haykins[1] defined Cognitive radio as an intelligent wireless communication system that is aware of and learn from it’s environment and adapts its internal states by making corresponding changes in certain operating parameters. In similar track many research had been done in reconfigurablity in parameters from definition of radio parameter, physical layer protocol change , modulation technique adaptation, MAC layer adaptation and in some place Network layer adaptation. Little research had been reported on cross layer aware protocols in wireless communication. Authors[4] in shows significantly

About[1]- Department of Information Technology, Suresh Gyan Vihar University, Jaipur, Indiavibhakarp@rediffmail.com
About[2]- Department of ECE, SBCET, Benad Road, Jaipur, India roy.krishna@rediffmail.com
About[3]- Department of Computer Engineering, Suresh Gyan Vihar University, Jaipur, India sksmtech@yahoo.com

improved protocol for MAC layer by adapting the cross layer aware system. The enhancement in performance is due to awareness and adaptation of PHY parameters. Authors[7] in shows significant improvement in routing protocol by using cross layer aware protocol design in network layer protocols. Noticeable improvement are also reported by authors[4] over KR-MAC (Knowledge based Reasoning MAC and CLA-AMAC .Routing in multi-hop heterogeneous wireless network using non adaptive routing system is not adequate because it selects minimum Hop count path ,which have significantly less capacity than the best paths that exist in the Network .

Bio inspired algorithm are mainly based on hybrid (both reactive and proactive) multipath algorithm, AntHocNet is one of the most respected bio inspired routing alogithm in MANET. In AntHocNet [8] a routing table consists of a destination,the next possible hop to it, and a special data structure based on odors released by ant called pheromone. A pheromoneis a value that indicates the estimated goodness of a path between a source and a destination. In this way, pheromone data structures in different nodes indicate multiple paths between two nodes in the network, and are stochastically spread over it (in each node they select the next hop with a probability proportional to its pheromone value). Once paths are set up and the data start to flow, the source node starts to send proactive forward ants to the destination. This is a maintenance phase where each proactive forward ant follows the pheromone values in the same ways as the data, but has a small probability at each node of being broadcast. This technique serves as follows. If the forward ant reaches the destination without a single broadcast, it means that the current path is working and optimal, and it provides an efficient way of data transferring. On the other hand, if the ant gets broadcast at any point, it leaves the currently known pheromone trails as knowledge base and it explores new paths. A threshold of value two (2) is used to avoid proactive forward ants being broadcast to the whole network, allowing the search for improvements or variations to be concentrated around the current paths. In the case of a link failure, a node may use an alternative path based on the pheromone values . However, if the failed link was the only one in each pheromone table, the node sends out a route repair ant that travels to the involved destination like a reactive forward ant would do. Simulation experiments have shown that AntHocNet can outperform AODV and other routing algorithm in terms of delivery ratio and average delay [8].

The layered architecture simplifies development of different components by keeping each layer isolated from the others. Originated from the wired networks world, the concept of transparency is what makes OSI, TCP/IP and IEEE 802 models allow rapid and universal development and improvements. Nevertheless, it has become evident that the traditional layered approach that separates routing, flow control, scheduling, and power control is suboptimal in the realm of wireless and agile networks. This can be attributed to the complex and unpredictable nature of the wireless medium. Thus, the need for adaptation in network protocols remains high. In order to tackle the problems faced in wireless agile networks, a cross layer design [9] is desired to optimize across multiple layers of the protocol stack. The basic idea of cross-layering is to make information produced or collected by a protocol available to the whole protocol stack, so as to enable optimization and improve network performance. Until now several approaches have been proposed by researchers that use cross-layering in order to improve and optimize different network mechanisms. In most of the cases, the cross-layer design takes place between the media access control (MAC) and the physical (PHY) layers[4]. However, there is a number of recent examples that illustrate the benefits of having other layers jointly designed, such as network-data link layer (DLL), or even application-network. For instance, in order to bypass the resource constraints, Shah and Rabaey [2] have proposed an energy-aware routing protocol that uses a set of suboptimal paths occasionally to increase the lifetime of the network. The idea is that paths are chosen by means of a probability and knowledgebase that depends on how low the energy consumption of each path is. The energy consumption is a result of signal strengths, a piece of knowledge that can be found at the MAC layer of the stack. Hence, cross-layering helped to access the information and use it to the network layer (routing layer) to make analogous decisions. Another example, this time in link-aware routing was proposed by Lee and Gerla [3]. This protocol makes use of channel state information (CSI) [6] and cross-layer integration to route traffic along higher-capacity paths by consistently selecting channels with favorable conditions. This supports the idea that a node with multiple next-hop alternatives can measure the channel state on the links, and then forward a packet based on the link quality and other metrics. Cross-layer has also been a great help in designing cost aware routing approaches. Suhonen et al [7] have proposed a protocol that uses cost metrics to create gradients from a source to a destination node. The cost metrics consist of energy, node load, delay, and link reliability information that provide traffic differentiation by allowing choice among delay, reliability, and energy.

III. Proposed protocol

This is a extension of cross layer aware protocol develop by authors [7] . In this information about channel state , observed link state and hop by hop reasoned and observed information are utilized by network layer protocol in general and routing algorithm in specific .Exactly Signal to interference and noise ratio(SINR) , received power(RP) , delay observed by reactive ant, pheromone value, knowledge based interpolation are passed on to routing algorithm for decisions for source routing between source and destination . The protocol improves over another cross layer protocol by employing ANT colonization approach for optimization and knowledge based reasoning for decision support .Apart from decision in proactive routing it can also adapt as per reconfigurability of PHY or flexibility provided by agile radio.

The protocol is based on source routing with additional information and decision parameters from PHY and MAC Layer. The protocol has very simple two fold approach. First fold use to discover the route ,which is as.

1) A small packet known as ANT is sent to discover new route.
2) ANT places small amount of data containing PHY and MAC observed information on to every node it traverse. It is just like ANT leaving pheromone in the route.
3) If ANT found destination route without broadcast. It can be thought as optimal route.
4) If ANT stuck at any node it broadcast with threshold 2 which guarantee non flooding of network and producing sub optimal alternate route.

In second fold the pheromone placed at each node is used by reasoning engine for short term prediction on link state and route condition in hop by hop basis . which is use to adapt optimize various communication parameter based on AgileMAC protocol develop by authors [4] .

IV. Simulation Results

The proposed cross-layer protocol has been implemented in theOMNET++ 4.0 network simulator [5]. The simulations have been carried out for various topologies, scenarios with different kinds of traffic, and routing protocols. The following performance metrics have been used:

(i) total packets received,
(ii) average throughput (Mbps),
(iii) lifetime LND (seconds),
(iv) FND: first active node died (seconds),
(v) lifetime RCVD (seconds),
(vi) average aggregate delay (seconds),

The first node died metric is defined as the instant in time when the active (a node transmitting/receiving) first node died. We have defined the network lifetime as the time duration from the beginning of the simulation until the instant when the active (a node transmitting/receiving) last node died, that is, there is no live transmitter-receiver pair left in the network. The Lifetime RCVD is specified as the instant in time when the last packet is received.

The average throughput has been defined as

Thr = Total numberPackets received Simulation Time [Mbps] and

average sending bit rate has been defined as

Sbit = Total numberPackets sent Simulation Time [Mbps].

Number of active nodes	25, 50 *(default)*
Simulations area	≤ 1000 * 1000m
Topology	Random
PHY/MAC	DSSS, IEEE 802.11b
SINR thr. (dB)	22.05
Type of netwok	*homo/hetero*-geneous
Initial energy (J) variable =	0.5–. . ., 5, 20
PtMAX −	250m 0.200888W
PtMAX −	100m 0.010072W
txPowerinit	250 _ 100 meters
rxPower	45% of PtMAX
idlePower	30% of PtMAX
Capture Thr.(dB)	10
Traffic model	CBR/UDP
Payload size (bytes)	2048 _ 100–8192
CWmin −CWmax (slots)	15–1023
Simulation time (s)	≤650
Movement	random and constant
Mobility model	turtle Model
Speed (m/s)	0 − 2 ≤ 20;
1.5− *(default)*	
Access scheme Basic	*(default)* _ RTS/CTS

Table 2: Typical values of path loss exponent and shadowing deviation.

Environment	ρ (dB)	σ (dB)
Outdoor Free space	2.4	4 to 12
Outdoor Shadowed Urban	2.7 to 5.6	4 to 12
Indoor Line-of-sight	1.6 to 1.8	3 to 6
Indoor Obstructed	4 to 6	6.8

We study the performance of the routing algorithms in different regimes. The performance metrics for the stationary and turtle mobility scenario in various offered load regimes are plotted and discussed as follows.

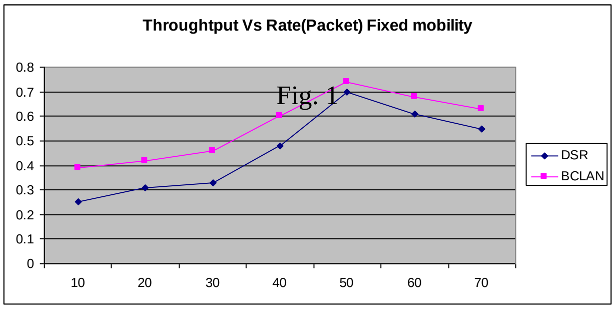


Fig. 1

Consider The Fixed mobility scenario in Fig. 1. As the offered load increase the average throughput increase, for DSR scheme congestion begins to buildup at packet rate of 40 packets/ sec.. The throughput falls slightly when packet rate is further increase. Since rate adaptation is used in BCLAN schemes ,the network capacity is much higher and congestion is non observant. The BCLAN offered 45% more throughput enhancement compared to the DSR scheme. The same effect with some modification can be seen in Fig 2. for turtule mobility

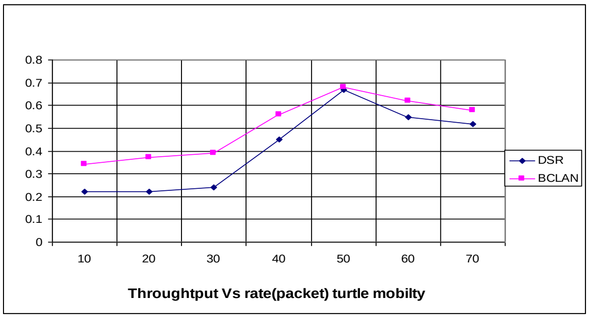


Fig. 2.

V. Conclusion

In this paper we advocate a new design concept in routing protocol based on bio inspired computing with prediction capability of reasoning engine. We argue that by exploiting information from MAC and PHY layer, significant performance enhancement of a routing protocol could be achieved. The proposed protocol BCLAN protocol produces 45% better result than the DSR algorithm .The can be used for broadbad services like VoD, VoIP.

VI. Reference

1) S. Haykin, "Cognitive Radio: Brain-empowered Wireless Communications," *IEEE J. Selected Areas in Communications*, **23**, No.2, pp.201- 220, Feb. 2005.
2) R. C. Shah and J. Rabaey, *Energy aware routing for low energy ad hocsensor networks*, pp. 350–355. In Proceeding of the IEEE Wireless Communications and Networking Conference, Orlando, FL, 2002.
3) S. J. Lee and M. Gerla, *AODV-BR: Backup routing in ad hoc networks*,pp. 1311–1316. In Proceeding of the IEEE Wireless Communicationsand Networking Conference, Chicago, 2000.
4) Vibhakar Pathak, K C Roy , Santosh K Singh , "Cross layer aware adaptive MAC based on knowledge based reasoning for cognitive radio computer network " in International Journal of Next-Generation Networks Vol. 2,No.2,June2010 Pg14-21
5) www.omnetpp.org
6) A. I. Perez-Neira and M. R. Campalans, *Cross-Layer Resource Allocationin Wireless Communications*. Elsevier, 2009.
7) J. Suhonen, M. Kuorilehto, M. Jannikainen, and T. D. Hamalainen, *Cost-aware dynamic routing protocol for wireless sensor networks design and prototype experiments*, pp. 1–5. In Proceeding of the 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Helsinki, Finland, 2006.
8) G. D. Caro, F. Bucatella, and L. M. Gambardella, Special Issue on Selforganisationin Mobile Networking: AntHocNet: an adaptive natureinsipredalgorithm for routing in mobile ad hoc networks. AEIT, 2005.
9) M. Ibnkahla, ed., Adaptation and Cross Layer Design in Wireless Networks. CRC Press, 2009.

Global Journals Inc. (US) Guidelines Handbook 2011

www.GlobalJournals.org

FELLOW OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (FICCT)

- FICCT' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FICCT" can be added to name in the following manner e.g. **Dr. Andrew Knoll, Ph.D., FICCT, Er. Pettor Jone, M.E., FICCT**
- FICCT can submit two papers every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free unlimited Web-space will be allotted to 'FICCT 'along with subDomain to contribute and partake in our activities.
- A professional email address will be allotted free with unlimited email space.
- FICCT will be authorized to receive e-Journals - GJCST for the Lifetime.
- FICCT will be exempted from the registration fees of Seminar/Symposium/Conference/Workshop conducted internationally of GJCST (FREE of Charge).
- FICCT will be an Honorable Guest of any gathering hold.

ASSOCIATE OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (AICCT)

- AICCT title will be awarded to the person/institution after approval of Editor-in-Chef and Editorial Board. The title 'AICCTcan be added to name in the following manner:
  eg. **Dr. Thomas Herry, Ph.D., AICCT**
- AICCT can submit one paper every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free 2GB Web-space will be allotted to 'FICCT' along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted with free 1GB email space.
- AICCT will be authorized to receive e-Journal GJCST for lifetime.
- A professional email address will be allotted with free 1GB email space.
- AICHSS will be authorized to receive e-Journal GJHSS for lifetime.

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJCST for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.

Online Submission: There are three ways to submit your paper:

(A) (I) Register yourself using top right corner of Home page then Login from same place twice. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal from "Research Journals" Menu.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer (Although Mozilla Firefox is preferred), then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org as an attachment.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION **(Must be strictly followed)**

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Times New Roman.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be two lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.

Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads: Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also. Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:
Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.
Abstract, used in Original Papers and Reviews:
Optimizing Abstract for Search Engines
Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words
A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.
One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

Informal Tips for writing a Computer Science Research Paper to increase readability and citation

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

Techniques for writing a good quality Computer Science Research Paper:

1. Choosing the topic- In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. **Pick a good study spot:** To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. **Think technically:** Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page

- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.

- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before
Submitting Your Research Paper to Global Journals Inc. (US)

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US) Inc.(US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

9 770116 586982 61427>

ISSN 9754350